1989

Heterogeneity of variances by herd production level and its effect on dairy cow and sire evaluation

Keith George Boldman
Iowa State University

INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book. These are also available as one exposure on a standard 35mm slide or as a 17" x 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 8920113

Heterogeneity of variances by herd production level and its effect on dairy cow and sire evaluation

Boldman, Keith George, Ph.D.

Iowa State University, 1989

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

Heterogeneity of variances by herd production level and its effect on dairy cow and sire evaluation

by

Keith George Boldman

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of the
Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Department: Animal Science
Major: Animal Breeding

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1989

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Genetic progress in a dairy cattle population is largely determined by the genetic superiority of the bulls selected by artificial insemination (AI) studs. Selection of young bulls is a two-stage procedure. In stage one, elite cows are mated to superior sires to produce young bulls. In the second stage, the young bulls are progeny tested and a small percentage are selected for extensive use throughout the cow population. Selection of inferior sires and dams to produce young bulls or inaccurate progeny testing of these young bulls will decrease realized genetic gain. Van Vleck (1977) reported that genetic progress in dairy cattle has been much less than theoretically possible.

Heterogeneous within-herd variances may decrease the accuracy of genetic evaluations of dairy cattle. Generally, the mean and variance of milk production is positively correlated so the differences between cows are greater in high yield herds. Most mixed model genetic evaluation procedures for dairy cattle, however, assume variances and heritability are equal for all records.

The presence of unequal variances presents problems in the selection of bull dams and the evaluation of young bulls. Potential bull dams from different herds are compared but the variance may differ from one herd to another. Cows with extreme levels of production relative to herdmates are more

likely to be found in herds with high variance. As a result, superior cows in herds with high variance will tend to be overevaluated relative to superior cows in herds with low variance. An apparent excess of elite cows has been reported for high variance herds (Everett et al., 1982) and high yield herds (Powell et al., 1983). Sires are progeny tested across many herds but the distribution of daughters may not be random with respect to within-herd variance. Sires with a large percentage of their daughters in high variance herds may be overevaluated. In the presence of heterogeneous variance, the estimated breeding value for an individual may be a function of herd variance as well as genotype and animals may be misranked.

Previous studies of heterogeneous variance have estimated variances of first lactation milk yield by production level using a sire model. Most studies have found that estimates of genetic and residual variances and heritability increase with production level. A simple method of equalizing variances is to transform data. Logarithmic transformation of yields has been proposed as a method to equalize variances across herds. Log transformation is appropriate for linear models in which group variances are proportional to group means. In order for log transformation to be effective for mixed models, the transformation should stabilize both genetic and residual variances.

Use of all lactations in the mixed model equations for cow and sire evaluation requires estimates of additive genetic, permanent environmental, and residual variances. The first objective of this study was to estimate these variances for untransformed and natural logarithmic transformed milk and fat yield at three levels of production. The second objective was to compare the genetic evaluations of dairy sires and cows from a mixed model accounting for heterogeneous variances across three production levels with evaluations from a model in which variances were assumed homogeneous across production levels.

SECTION I. VARIANCE COMPONENT ESTIMATION

Abstract

Additive genetic, permanent environmental, and residual components of variance were estimated by restricted maximum likelihood for all lactation milk and fat yields and natural logarithm of yields at three production levels. Data consisted of 121,136 mature equivalent, 2x, 305-d first and later lactation yields for 91,206 Holstein cows calving between 1979 and 1984 throughout the United States. A total of 526 sires were represented and 485 of these had first crop daughters in the data. Production levels were defined by mean mature equivalent milk yield of all cows freshening in the same herd-year. Model of analysis included fixed herd-year-season and sire genetic group and random sire nested within group, cow nested within sire, and residual effects.

For untransformed milk yields, variance components increased with production level, and heritability and repeatability were smallest for low yield herds. After log transformation, permanent environmental and residual variance decreased as production level increased and genetic variance increased at a smaller relative rate. Log transformation did not change heritability and repeatability of milk yield.

Variance estimates of untransformed fat yield increased with production level for all components except permanent environment which was equal for medium and high herds. For

log transformed fat yield, residual variance decreased as production level increased, and genetic variance was largest and permanent environmental variance was smallest for low-production herds. Log transformation did not change heritability or repeatability estimates which increased with production level. Correlations of sire solutions across production levels were close to expected values for both milk and fat yield, indicating the absence of genotype-environment interaction.

These results indicate that it may be necessary to account for heterogeneous variances of all lactation yields in sire and cow evaluation. A single transformation will not equalize all variance components across all production levels.

Introduction

Most mixed model genetic evaluation procedures for dairy cattle assume genetic and residual variances are equal for all records. This simplifying assumption may be incorrect if variance of milk yield is greater in high-producing herds (Hill et al., 1983). If the relationship between mean and variance is a scale effect, i.e., caused by the scale of measurement, it can often be removed by an appropriate transformation (Falconer, 1981).

Logarithmic transformation of yields has been proposed as a method to equalize variances across herds. Everett et al. (1983) estimated residual variances for individual herds by

summing squared residuals from a model used to estimate cow breeding values in the northeastern United States for milk yield in 1981. For untransformed data, the estimated regression of within-herd error standard deviation on average herd production was +.083 kgs, i.e., herd variance increased with herd average. After natural log transformation of yields, the estimated regression was -.025. Estimates for genetic variance and heritability after transformation were not presented. The researchers concluded that log transformation removed a large part of the association between mean and variance of milk yield. Log transformed production records are currently used to equalize error variances in the Northeast AI Sire Comparison (Everett et al., 1983).

Several studies have estimated variance components of untransformed and log transformed milk yield from herd-year-seasons (HYS) grouped by production level. Hill et al. (1983) estimated variance components and heritability of first lactation milk yield from daughters of British Friesian sires. Herds were split into two production levels on the basis of the mean milk yield of all first lactation cows. Estimates of sire and residual variance and heritability for untransformed records were greater for the high-production level. For the logarithm of records, residual variance was essentially equal for both levels and the increase in sire component was smaller than on the original scale. The increase in heritability was

greater on the log scale. The correlation between sire evaluations in low and high yield herds was close to unity, indicating the absence of genotype-environment interaction.

Mirande and Van Vleck (1985) estimated heritability of milk yield at four production levels by a sire model. Data were first lactation mature equivalent (ME) milk records of artificially sired Holstein cows in the northeastern United States. Rolling herd average was used to assign cows to production level, and variance components were estimated by year of freshening. Heritability estimates were averaged over years by weighting by number of records in each year. On the original scale, sire and residual components of variance increased with production level. On the log scale, however, residual variance was largest at low-production and smallest at high-production levels. Transformation of yields did not change estimates of heritability which were highest for middle-production and smallest for low-production groups.

Variance components of milk yield were estimated by De Veer and Van Vleck (1987) using a multiple trait sire model in which yields at three production levels were considered correlated variables. Records used were first lactation, ME records of Holstein cows sired by AI bulls in the northeastern United States. Analyses were performed for each of four years separately. Cows were assigned to production level on the basis of mean yield of all cows freshening in the same HYS.

For untransformed records, estimates of sire and residual variance components and heritabilities within years increased as the mean increased. Estimates of genetic correlations among expressions of genotype in low, medium, and high-production herds were close to unity. After logarithmic transformation of yields, sire components of variance were essentially equal but residual components decreased as the production level increased. Heritabilities for log transformed records increased with production level and were larger than for untransformed records.

These and most other studies have indicated a positive relationship between production level and estimates of genetic and residual variances and heritabilities. Heritability estimates of milk yield seem to be heterogeneous whether records are expressed on untransformed or log scales.

Previous studies have estimated variances and heritabilities of first lactation yield. An all lactation animal model has recently been developed for genetic evaluation of dairy cows and sires in the United States (Wiggans et al. 1988a, 1988b). Use of all lactations in a mixed model requires estimates of additive genetic, permanent environmental, and residual variances. The objective of this study was to estimate these variances for milk and fat yields on the untransformed and natural logarithmic scale at three levels of production.

Materials and Methods

Description of data

Data were provided by the United States Department of Agriculture Animal Improvement Programs Laboratory (USDA-AIPL) and consisted of mature equivalent (ME), 2x, 305-d milk and fat yields from first and later lactations of artificially sired Holstein cows throughout the United States. Cows less than 19 months or greater than 36 months of age at first parity were eliminated. Records with less than 1361 or greater than 18,144 kg ME milk, or less than 50 or greater than 635 kg ME fat were eliminated. Each record was required to have at least 60 days in milk (DIM). The DIM requirement used was less than in most previous studies but Aisbett (1984) has suggested that association of herd means and variances may in part be due to edits for minimum lactation length. Two seasons, November through April and May through October, were defined per year.

Data included sampling and later daughters and herdmates of 485 sires born between 1975 and 1978. Records on which selection was based must be included in an analysis in order for estimates of variance components to be free from selection bias (Henderson, 1984b). The original intent was to include only sires with sampling daughters to reduce the effect of selection on sire variance. Inclusion of only sires with sampling daughters in the analysis, however, would have

resulted in a large number of cows without a contemporary for comparison. Therefore, daughters of 41 proven sires born in 1973 and 1974 were added to increase comparisons within HYS. Inclusion of proven sires would bias sire variance estimates downward (Robertson, 1977), but the objective was to compare variance estimates in different production levels. Because the proven sires were included in the analysis at each production level, selection was not thought to significantly affect the comparison of variances across production levels.

The inverse numerator relationship matrix due to sires and maternal grandsires was computed by the method of Henderson (1975). A total of 92 common sires and maternal grandsires of the 526 sires with daughters was identified; all but three of the sires with daughters had a sire and/or maternal grandsire in common. These additional sire and maternal grandsire relationships were absorbed into those for sires with daughters resulting in an inverse relationship matrix of order 526.

Definition of production levels

Each record in the analysis was assigned to one of three production levels on the basis of mean ME milk yield of all cows freshening in the same herd-year, regardless of whether the cow was included in the analysis or not. Upper and lower limits used to define production levels were: low, 5897-7484 kg; medium, 7485-8618 kg; and high, 8619-10,206 kg. The mid-

point of the medium level was approximately equal to the mean for all records and limits for the medium level covered a smaller range to increase number of records at low and high levels. Records from herd-years with averages less than 5897 or greater than 10,206 kg were eliminated. Each cow was required to have a first lactation record to avoid selection bias due to culling. Later lactations for a cow were included if they were assigned to the same production level as the first lactation. Therefore, each cow was included in the analysis for only one production level while most sires were represented by cows at all three levels. After elimination of records coming from HYS represented by only one sire, data consisted of 121,136 first and later lactations initiated between November 1979 and April 1984 by 91,026 cows out of 526 sires. Number of lactations per cow ranged from one to five. Approximately 75% of the cows at each production level had only a first lactation in the analysis. Table 1 contains the number of HYS, sires, cows, and lactations for each production level. Number of cows and records per HYS increased across production levels indicating a positive relationship between herd size and herd average.

Table 1. Number of herd-year-seasons, sires, cows, and lactations at each level of production

Level of production	Herd-year-seasons	Sires	Cows	Lactations
Low	5801	515	18,754	24,659
Medium	10,952	526	43,103	57,895
High	5772	516	29,169	38,582

Model and estimation procedure

The twelve combinations of yield (ME milk or ME fat), production level (low, medium, or high), and scale (untransformed or log transformed) were analyzed separately. The assumed model for estimation of variance components included fixed herd-year-season (h) and sire genetic group (g) effects and random sire (s) nested within group, cow (c) nested within sire, and residual (e) effects:

$$y_{ijkln} = h_i + g_j + s_{jk} + c_{jkl} + e_{ijkln} \quad [1]$$

where y_{ijkln} is ME yield of record n of cow l of sire k in genetic group j in herd-year-season i. Sires were assigned to one of six genetic groups by year of birth. Number of sires assigned to birth year groups 1973 through 1978 were 22, 19, 146, 165, 138, and 36, respectively. The model in [1] can be expressed in matrix notation as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zs} + \mathbf{Wc} + \mathbf{e} \quad [2]$$

with

$$E\begin{bmatrix} y \\ s \\ c \\ e \end{bmatrix} = \begin{bmatrix} Xb \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and}$$

$$\text{Var}\begin{bmatrix} y \\ s \\ c \\ e \end{bmatrix} = \begin{bmatrix} ZAZ'\sigma_s^2+WW'\sigma_c^2+I\sigma_e^2 & ZA\sigma_s^2 & W\sigma_c^2 & I\sigma_e^2 \\ & A\sigma_s^2 & 0 & 0 \\ \text{symmetric} & & I\sigma_c^2 & 0 \\ & & & I\sigma_e^2 \end{bmatrix}$$

where y, b, s, c, and e are vectors of yields, fixed effects (herd-year-seasons and genetic groups), sires, cows, and residuals, respectively; **X**, **Z**, and **W** are design matrices for b, s, and c, respectively; **A** is a matrix of additive relationships among sires; and σ_s^2, σ_c^2, and σ_e^2 are variances of s, c, and e effects, respectively. Cows were assumed unrelated to simplify computations. The mixed model equations (MME) for model [2] are:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z+A^{-1}\alpha_s & Z'W \\ W'X & W'Z & W'W+I\alpha_c \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{s} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix} \qquad [3]$$

where $\alpha_s=\hat{\sigma}_e^2/\hat{\sigma}_s^2$ and $\alpha_c=\hat{\sigma}_e^2/\hat{\sigma}_c^2$.

Restricted maximum likelihood (REML) estimation of variance components in a sire and nested cow model has been

shown to substantially reduce biases due to cow culling (Ouweltjes et al., 1988). REML method estimates of variance components using an expectation maximization (EM)-like algorithm were obtained through an iterative scheme using the MME. Let a generalized inverse of the coefficient matrix in [3] be:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\alpha_s & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'W}+\mathbf{I}\alpha_c \end{bmatrix}^{-} = \begin{bmatrix} C_{bb} & C_{bs} & C_{bc} \\ C_{sb} & C_{ss} & C_{sc} \\ C_{cb} & C_{cs} & C_{cc} \end{bmatrix} . \quad [4]$$

Variance components were estimated in each round of iteration from (Meyer, 1987):

$$\hat{\sigma}_s^2 = \hat{s}'\mathbf{A}^{-1}\hat{s} \,/[ns - \alpha_s \text{trace}(\mathbf{A}^{-1}C_{ss})] \,, \quad [5]$$

$$\hat{\sigma}_c^2 = \hat{c}'\hat{c} \,/[nc - \alpha_c \text{trace}(C_{cc})] \text{ , and} \quad [6]$$

$$\hat{\sigma}_e^2 = \hat{e}'\hat{e} \,/[ndfe + \alpha_s \text{trace}(\mathbf{A}^{-1}C_{ss}) + \alpha_c \text{trace}(C_{cc})] \,. \quad [7]$$

where n, ns, and nc are the number of observations, sires, and cows, respectively; ndfe=n-ns-nc-rank(**X**), the degrees of freedom for error; and $\hat{e}'\hat{e}=y'(y-\mathbf{X}\hat{b}-\mathbf{Z}\hat{s}-\mathbf{W}\hat{c})-\alpha_s\hat{s}'\mathbf{A}^{-1}\hat{s}-\alpha_c\hat{c}'\hat{c}$. A computing algorithm for estimating variance components by REML for a model with two random effects and a large number of fixed effect subclasses has been described by Meyer (1987). In this procedure, cows and fixed effects are absorbed into sires and the trace of the inverse corresponding to cows is determined indirectly (see Appendix A for a description and numerical example). Convergence was assumed when the change

in each of the estimates from the previous round was less than .1%.

Computational aspects

The EM algorithm yields non-negative estimates (Harville, 1977) but is often slow to converge, especially when the residual variance is large relative to other variances (Laird and Ware, 1982), i.e., when heritability and repeatability are small. Several approaches have been suggested to reduce computational demands of the EM algorithm. Most of these methods can be classified into one of two general categories: 1) methods to reduce the number of rounds required to reach convergence, or 2) methods to reduce the cost per round of iteration. In this study, several methods were utilized in an attempt to reduce computational costs.

Methods to speed convergence Thompson and Meyer (1986) described a reparameterization to speed the convergence rate of the EM algorithm. The reparameterization was derived considering the expectations of mean squares in a balanced analysis of variance. When the reparameterization was applied to the data in the analysis, however, negative estimates were obtained after only two or three rounds of iteration. Thompson and Meyer (1986) warned that the reparameterization can lead to negative estimates, especially if true values of variances are close to zero. This explanation does not seem plausible in this analysis because all estimates of variance

components obtained without the reparameterization were greater than zero. A possible reason for negative estimates was the model used for analysis. The reparameterization was originally derived for a model in which the second random factor, e.g., cows, is nested within the first, e.g., sires. In the model used, however, the first random factor, i.e., sires, was also nested within a fixed effect, i.e., genetic groups. As a result, expectations of mean squares in an analysis of variance would be different and a different form of the reparameterization might be required.

VanRaden and Freeman (1987) presented a REML algorithm denoted EMC. In the EMC algorithm, a term approximating the actual expectation of residual variance is subtracted from each quadratic. The algorithm was originally developed for a sire model, but it has been applied to a model with two random effects in a nested design (Buttram, 1987) resulting in a reported 25% decrease in the rounds of iteration required to reach convergence. Use of the EMC algorithm did not speed convergence, however, for the data and model used in the study.

Neither of these methods decreased number of rounds required to reach convergence. Rounds of iteration ranged from 20 to 42, but from the middle to final round of iteration, heritability and repeatability estimates changed only at the fourth significant digit.

Methods to reduce cost per round Inversion of the absorbed sire coefficient matrix was the most computationally demanding step in each round of iteration. Vectorization is a process in which groups of numbers are processed as a unit with a single instruction. DO loops in the inversion subroutine were converted into array operations (VAST-E User's Guide, 1988) to increase processing speed. Central processing unit (CPU) time required to invert a matrix is a function of the order of the matrix cubed. Ninety-two common sires and maternal grandsires of the 526 bulls with daughters were identified resulting in a sire coefficient matrix of order 618. To reduce the order of the matrix, relationships for sires and maternal grandsires were absorbed, resulting in a matrix of order 526 to be inverted. The CPU time required to set up the full $\mathbf{A}^{-1}$ and absorb sires and maternal grandsires was less than 1 minute on a National Advanced Systems AS/9160 computer. After absorption, CPU time required for inversion in each round was reduced by over one-third:

$$\frac{\text{sire, mgs absorbed } \mathbf{A}^{-1}}{\text{full } \mathbf{A}^{-1}} \quad \frac{(526)^3}{(618)^3} = .62 \ .$$

In addition to reducing the cost per round, absorption of sires without progeny has been shown to reduce the number of rounds required to reach convergence (see Appendix B for a proof of the equivalence of the alternative sire quadratics and a numerical example).

Estimation of genetic parameters

Expectations of sire and cow components of variance were:

$$E(\hat{\sigma}_s^2) = 1/4\sigma_a^2$$

and

$$E(\hat{\sigma}_c^2) = 3/4\sigma_a^2+\sigma_p^2$$

where σ_a^2 is additive genetic variance of yield and σ_p^2 is variance of nonadditive genetic effects and permanent environmental effects. Based on these expectations, σ_a^2 was estimated as $4*\hat{\sigma}_s^2$ and σ_p^2 as $\hat{\sigma}_c^2-(3*\hat{\sigma}_s^2)$. Estimates of variance components were used to estimate heritability and repeatability of yield as:

$$\hat{h}^2 = \hat{\sigma}_a^2/(\hat{\sigma}_a^2+\hat{\sigma}_p^2+\hat{\sigma}_e^2)$$

and

$$\hat{r} = (\hat{\sigma}_a^2+\hat{\sigma}_p^2)/(\hat{\sigma}_a^2+\hat{\sigma}_p^2+\hat{\sigma}_e^2) \ .$$

The variance component estimation procedure using the MME also produced best linear unbiased procedure predictions of sire transmitting ability ($\hat{g}_j+\hat{s}_{jk}$) at the three production levels. Falconer (1952) suggested that genotype-environment interaction could be measured by the correlation between progeny performance in two environments. A low correlation across production levels indicates the presence of genotype-environment interaction. Estimates of product-moment correlations were calculated between predictions of transmitting ability for the 485 sires with first crop daughters at different herd production levels. The ratio of

estimated (r_{est}) and expected (r_{exp}) correlations was used to estimate genetic correlations between production levels q and q' from (Calo et al., 1973):

$$\hat{r}_{g_q g_{q'}} = r_{est}/r_{exp} \cdot \qquad [8]$$

Expected correlations of progeny tests were calculated as a function of accuracy of sire transmitting abilities (Hickman et al., 1969):

$$\hat{r}_{exp} = (\Sigma_i r^2_{s\hat{s}_{iq}} r^2_{s\hat{s}_{iq'}})/(\Sigma_i r^2_{s\hat{s}_{iq}} \Sigma_i r^2_{s\hat{s}_{iq'}})^{.5} \qquad [9]$$

where $r^2_{s\hat{s}_{iq}}$ is the square of the correlation between the true and predicted transmitting ability of sire i for trait q, i.e., squared accuracy of prediction, and the summation is over all sires. Expected correlations are dependent upon differences in the accuracy of sire transmitting ability predictions at each level. If a sire has many progeny in each environment, $r^2_{s\hat{s}}$ approaches one and the expected correlation in [9] approaches one. Danell (1982) approximated this accuracy for a sire as $r^2_{s\hat{s}}=n^*/(n^*+\hat{\sigma}^2_e/\hat{\sigma}^2_s)$ where n^* is effective number of progeny, defined as a diagonal element of the absorbed coefficient matrix for sires ignoring relationships. This expression underestimates accuracy if relationships among sires are included, and as a result, the expected correlation is underestimated. Sire accuracy can be expressed as a direct function of prediction error variance (PEV) from:

$$r^2_{s\hat{s}} = 1 - PEV/\sigma^2_s \cdot \qquad [10]$$

Unlike effective number of progeny, PEV accounts for additional accuracy resulting from the use of relationships among sires. PEV can be found from the appropriate sire diagonal element of the inverse of the MME. Because relationships among sires were considered in the analysis, the estimate of accuracy used in [9] was:

$$r^2_{s\hat{s}} = 1 - C_{ss}(\hat{\sigma}^2_e/\hat{\sigma}^2_s) \qquad [11]$$

where C_{ss} is a diagonal element of the inverse of the sire coefficient matrix in [4], and the residual and sire variances were estimated from [5] and [7], respectively.

Results

Means of records by production level for untransformed and log transformed milk and fat yields are in Table 2. Increase in mean from low to medium level was greater than the increase from medium to high-production level for all data sets.

Table 2. Means of untransformed and log transformed milk and fat yield by level of production

Level of production	Mean milk yield		Mean fat yield	
	linear kg	1000 * log kg	linear kg	1000 * log kg
Low	6954	8829	253	5515
Medium	8073	8981	292	5661
High	9180	9111	327	5776

Milk yield

Estimated variance components for untransformed and log transformed milk yields at the three production levels are listed in Table 3. To facilitate a comparison on the linear

and log scales, variance estimates are plotted as a percentage of the largest residual variance on each scale in Figure 1. Estimates for each variance component increased with production level on the untransformed scale. Genetic variance increased at the greatest relative rate, doubling from the low to high-production level. The relative increase in genetic variance with mean yield was smaller on the log scale as compared to the untransformed scale. Permanent environmental and residual variance decreased as production level increased after log transformation, resulting in the largest estimates at the low level.

Table 3. Estimated variance components of untransformed and log transformed milk yield by level of production

Level of production	Untransformed yield			Log yield		
	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_e$	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_e$
	(kg^2)			(1000 * kg)^2		
Low	258,203	483,789	731,731	5539	11,006	17,658
Medium	386,705	550,655	805,389	6176	9504	14,087
High	510,055	634,780	958,277	6490	8728	12,770

Estimates of heritability and repeatability for milk yield are in Table 4. Estimates averaged over production levels were similar to the values of .20 and .50 for heritability and repeatability, respectively, used in the USDA national animal model dairy evaluation (Wiggans et al., 1988a, 1988b). Heritabililites and repeatabilities were slightly less for low average yield herds and similar for medium and high herds. At

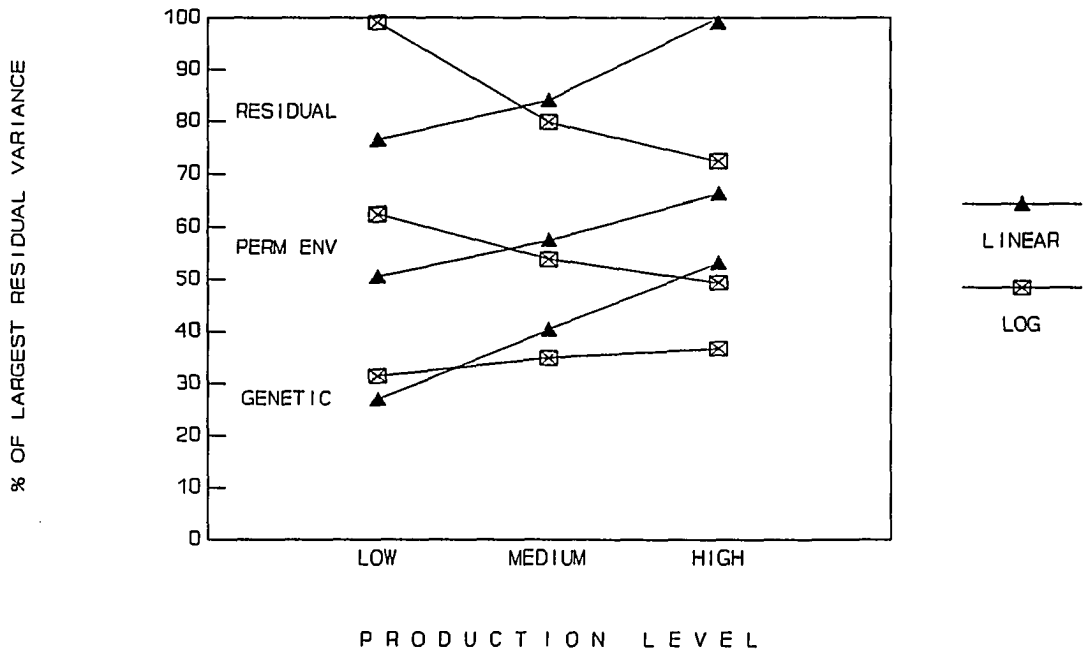

Figure 1. Variance components of milk yield relative to the largest residual variance by level of production on linear and logarithmic scales

Table 4. Estimated heritability and repeatability of untransformed and log transformed milk yield by level of production

Level of production	Untransformed yield		Log yield	
	$\hat{h}^2$	$\hat{r}$	$\hat{h}^2$	$\hat{r}$
Low	.18	.50	.16	.48
Medium	.22	.54	.21	.53
High	.24	.54	.23	.54

each production level, estimates were similar for both scales.

Genetic correlations between milk yield at the three production levels (Table 5) were close to unity, indicating the absence of genotype-environment interaction. The lowest

estimated correlation was between production in low and high herds. Genetic correlations were estimated as the ratio of estimated and expected correlations of sire solutions which resulted in estimates greater than 1.0. The expected correlation of sire solutions from equation [9] is sensitive to sire solutions based on few daughters, i.e., low accuracy (Blanchard et al., 1983). In these data, there was large variation in the accuracies of sire solutions across production levels for many sires.

Table 5. Genetic correlations between production in low, medium, and high average herds estimated from sire solutions for milk and fat yield

	Production levels compared		
Trait	Low to medium	Low to high	Medium to high
Milk yield	.99	.90	1.02
Fat yield	1.05	.96	1.11

Fat yield

Production levels for fat yield were also assigned on the basis of mean milk yield. Because milk and fat yield are not perfectly correlated, a better approach would be to assign levels by mean fat yield. Variance estimates for untransformed and log transformed fat yield at the three production levels are listed in Table 6 and plotted in Figure 2 as a percentage of largest residual variance. Estimates of variances for untransformed fat yield increased with production level for all components except permanent

environment which was equal for medium and high-production herds. The increase was greatest for genetic variance which doubled from the low to high level. Genetic variance and permanent environmental variance were homogeneous at low and medium levels after log transformation. On the log scale, genetic variance was largest at the high-production level, but permanent environmental variance was smallest at the high level. The pattern for residual variance was reversed after log transformation; estimates decreased as production level increased.

Table 6. Estimated variance components of untransformed and log transformed fat yield by level of production

Level of production	Untransformed yield			Log yield		
	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_e$	$\hat{\sigma}^2_a$	$\hat{\sigma}^2_p$	$\hat{\sigma}^2_e$
	———	(kg^2)	———	———	(1000 * kg)^2	———
Low	352	486	939	5993	7994	16,720
Medium	487	586	1056	6060	7771	13,706
High	741	588	1242	7538	6132	12,766

Heritability and repeatability for fat yield (Table 7) increased with production level on the linear and log scale; transformation did not change estimates of either parameter. Estimated genetic correlations between fat yield at the three levels (Table 5) were close to unity and two estimates were greater than 1.0. Equation [9] may underestimate expected correlation of sire evaluations, resulting in estimates of genetic correlation greater than 1.0.

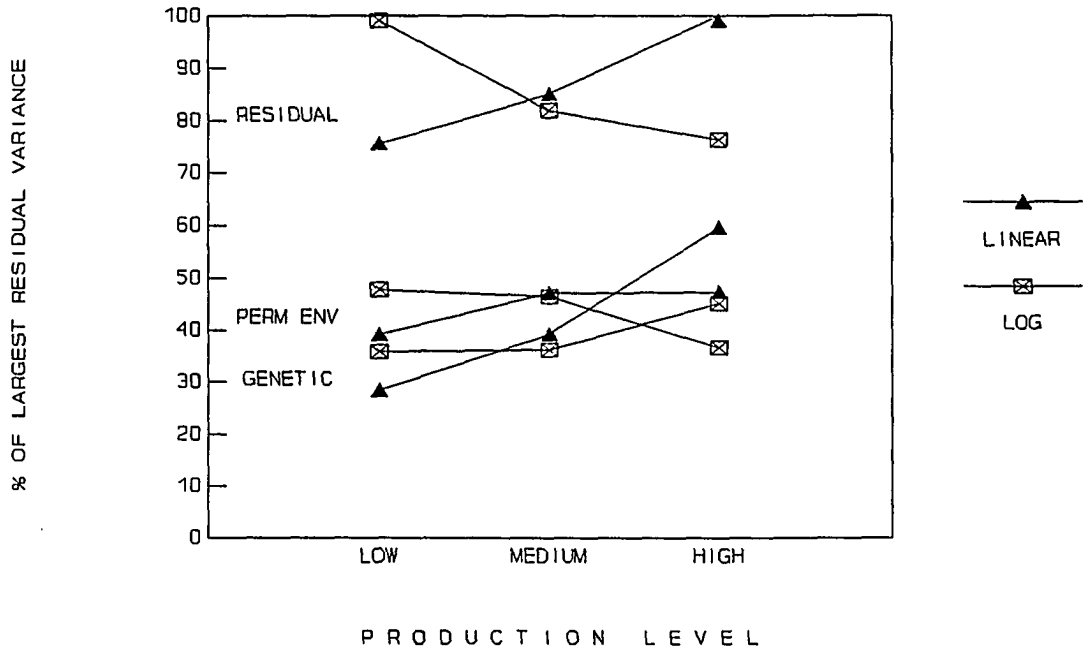

Figure 2. Variance components of fat yield relative to the largest residual variance by level of production on linear and logarithmic scales

Table 7. Estimated heritability and repeatability of untransformed and log transformed fat yield by level of production

Level of production	Untransformed yield		Log yield	
	$\hat{h}^2$	$\hat{r}$	$\hat{h}^2$	$\hat{r}$
Low	.20	.47	.20	.46
Medium	.23	.50	.22	.50
High	.29	.52	.29	.52

Discussion and Conclusions

The differences among variance components for all lactation milk and fat yields estimated by production level were large in these data. Use of equal variances for all

records in cow and sire evaluation is a simplifying assumption and is probably incorrect.

Consequences of using homogeneous variances need to be investigated. Powell et al. (1983) concluded that adjustments for larger variances and heritability in high producing herds would partially offset each other. For milk yields on the untransformed scale, lower heritability estimates at the low-production level would tend to offset the effect of smaller variances. Heritability and repeatability estimates for milk yield were similar at the medium and high levels but variances were larger at the high level. As a result, superior cows in high yield herds would be overevaluated relative to superior cows in medium herds. Variances of untransformed fat yield increased with production level, but estimates of heritability were similar for low and medium herds and greater for high level herds. Use of homogeneous fat yield variances would result in underevaluation of above average cows in low herds relative to cows in medium-production herds.

Genetic variances were less heterogeneous on the log scale. Log transformation of yields did not change estimates of heritability but residual variances of transformed records decreased across production level. Garrick and Van Vleck (1987) concluded from a simulation study that the greatest reduction in selection response results when heritability is greatest in the least variable environment. Log

transformation of yields may be a worse approach than simply ignoring heterogeneous variances.

The finding that logarithmic transformation does not adequately correct for heterogeneity of all variance components is not unexpected. In the case of a normally distributed variable X with mean μ and variance σ^2, the logarithm of X has approximate variance σ^2/μ^2 (Van Vleck, 1988). Therefore, log transformation will stabilize variances only if the standard deviation on the original scale varies directly with the mean, i.e., if coefficients of variation (CV) are equal. Hill et al. (1983) and Brotherstone and Hill (1986) have reported that CV of milk yield is heterogeneous across levels of production. For these data, CV of milk yield for low, medium, and high-production levels was .0731, .0770, and .0778 for genetic effects and .123, .111, and .107 for residual effects. Estimated variances on the log scale displayed the same trend; genetic variance increased slightly while residual variance decreased across production levels, and for both components the largest difference was between low and medium estimates. Because of the differences in CV, a single transformation is unlikely to stabilize genetic, permanent environmental, and residual variances across production levels. Falconer (1981) warned that different scales may be appropriate for genetic and environmental components of variation.

Several researchers have reviewed approaches for dealing with heterogeneous variances in mixed model evaluations. A multiple trait approach (Gianola, 1986; Henderson, 1984a), treating genetic value at each production level as a different trait, is not indicated because genetic correlations across environments have been found to be close to unity in this and other studies. A possible two-step procedure would be to first equalize genetic variance across environments and then use heterogeneous permanent environmental and residual variances in the analysis. In this study, log transformation of yields reduced but did not eliminate the heterogeneity of genetic variance. Genetic variance could be standardized by dividing each record by the appropriate genetic standard deviation (Gianola, 1986; Weller et al., 1985). Permanent environmental and residual variance matrices, heterogeneous by management group or production level, would then be used in the mixed model equations.

Rankings of sires and cows from a model assuming constant variances should be compared to a model accounting for heterogeneity of variance components. The assumption of homogeneous variances may result in a reduction in potential genetic gain.

References

Aisbett, C. W. 1984. Association of herd means and variances is a function of edit for minimum lactation length. J. Dairy Sci. 67:702.

Blanchard, P. J., R. W. Everett, and S. R. Searle. 1983. Estimation of genetic trends and correlations for Jersey cattle. J. Dairy Sci. 66:1947.

Brotherstone, S. and W. G. Hill. 1986. Heterogeneity of variance amongst herds for milk production. Anim. Prod. 42:297.

Buttram, S. T. 1987. Genetics of racing performance in the American Quarter Horse. Unpublished Ph.D. Dissertation. Iowa State University, Ames, IA.

Calo, L. L., R. E. McDowell, L. D. Van Vleck, and P. D. Miller. 1973. Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. J. Anim. Sci. 37:676.

Danell, B. 1982. Interaction between genotype and environment in sire evaluation for milk production. Acta Agric. Scand. 32:33.

De Veer, J. C. and L. D. Van Vleck. 1987. Genetic parameters for first lactation milk yield at three levels of herd production. J. Dairy Sci. 70:1434.

Everett, R. W., R. L. Quaas, and E. J. Pollak. 1983. Multiple trait Northeast artificial insemination sire evaluation. Anim. Sci. Mimeo Series No. 74. Cornell University, Ithaca, NY.

Falconer, D. S. 1952. The problem of environment and selection. Am. Nat. 86:293.

Falconer, D. S. 1981. Introduction to quantitative genetics. Longman, Inc., New York, NY.

Garrick, D. J. and L. D. Van Vleck. 1987. Aspects of selection for performance in several environments with heterogeneous variances. J. Anim. Sci. 65:409.

Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogeneous. Theor. Appl. Genet. 72:671.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statist. Assoc. 72:320.

Henderson, C. R. 1975. Inverse of a matrix of relationships due to sires and maternal grandsires. J. Dairy Sci. 58:1917.

Henderson, C. R. 1984a. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada.

Henderson, C. R. 1984b. Recent developments in variance and covariance estimation. J. Anim. Sci. 63:2082.

Hickman, C. G., A. J. Lee, and K. Gravier. 1969. Genotype x season x method interaction in evaluating dairy sires from progeny records. Can. J. Anim. Sci. 49:151.

Hill, W. G., M. R. Edwards, M. K. A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition at different levels of variability of production. Anim. Prod. 36:59.

Laird, N. M. and J. H. Ware. 1982. Random effects models for longitudinal data. Biometrics 38:963.

Meyer, K. 1987. Restricted maximum likelihood to estimate variance components for mixed models with two random factors. Génét. Sél. Evol. 19(1):49.

Mirande, S. L. and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.

Ouweltjes, W., L. R. Schaeffer, and B. W. Kennedy. 1988. Sensitivity of methods of variance component estimation to culling type of selection. J. Dairy Sci. 71:773.

Powell, R. L., H. D. Norman, and B. T. Weinland. 1983. Cow evaluation at different milk yields of herds. J. Dairy Sci. 66:148.

Robertson, A. 1977. The effect of selection on the estimation of genetic parameters. Z. Tierz. Zuchtungsbiol. 94:131.

Thompson, R. and K. Meyer. 1986. Estimation of variance components: What is missing in the EM algorithm? J. Statist. Comput. Simul. 24:215.

Van Vleck, L. D. 1988. Alternatives for evaluations with heterogeneous genetic and environmental variances. J. Dairy Sci. 71(Suppl. 2):83. (Abstr.)

VanRaden, P. M. and A. E. Freeman. 1987. Rates of convergence of REML algorithms. Iowa Agric. and Home Econ. Exp. Stn., Journal Paper No. J-12584.

VAST-E User's Guide. 1988. Edition 1.4. Pacific-Sierra Corp., Los Angeles, CA.

Weller, J. I., M. Ron., and R. Bar-Anan. 1985. Accounting for environmentally dependent variance components in BLUP sire evaluations. J. Dairy Sci. 68(Suppl. 1):212. (Abstr.)

Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988a. Animal model evaluation of Ayrshire milk yield with all lactations, herd-sire interaction, and groups based on unknown parents. J. Dairy Sci. 71:1319.

Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988b. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J. Dairy Sci. 71(Suppl. 2):54.

SECTION II. ADJUSTMENTS FOR HETEROGENOUS VARIANCES

Abstract

Genetic evaluations of dairy sires and cows from three alternative models were compared to determine if adjustment for heterogeneous variances resulted in important rank differences from a model assuming homogeneity of variances. Data consisted of 121,136 first and later lactation mature equivalent, 2x, 305-d milk yields for 91,026 Holstein cows out of 526 sires. The model for analysis included fixed herd-year-season and sire genetic group and random animal, permanent environmental, and residual effects. In two models, homogeneous variance components were used for all records and the analysis used untransformed or natural log transformed yields. In the third model, untransformed yields were analyzed and different variance components at three production levels were used in the mixed model equations. Each herd-year-season was assigned to the low, medium, or high-production level by the mean milk yield of all cows freshening in the same herd-year.

Rank correlations of all evaluations from the three models were greater than .99 for both sires and cows. Adjustment for heterogeneous variances had a greater effect on cow evaluation than sire evaluation. Differences in ranks of top cows were large across the three models. Both log transformation of yields and use of heterogeneous variances increased the

percentage of elite cows at the low and medium-production levels and decreased the percentage at the high level. If homogeneous variances are assumed in an evaluation model, use of untransformed yields is recommended over use of log transformed yields. A model accounting for heterogeneous variances at several production levels resulted in important changes in cow evaluation and is computationally feasible if variance component estimates are available.

Introduction

Several studies have estimated variance components of milk yields from herds grouped by production level (e.g., De Veer and Van Vleck, 1987; Hill et al., 1983; Mirande and Van Vleck, 1985). These studies have indicated a positive relationship between production level and estimates of variances, both genetic and residual. Most mixed model genetic evaluation procedures for dairy cattle use the simplifying assumption of equal genetic and residual variances for all records. If variances increase with mean yield but are assumed to be homogeneous, animals could be misranked. Superior cows in herds with large variances or sires with a large percentage of their daughters in large variance herds would tend to be overevaluated. Powell et al. (1983) found an excess of elite cows for high yield herds from an analysis assuming equal genetic and residual variances for all records.

Several researchers have presented approaches to adjust for heterogenous variances in mixed model evaluations. Van Vleck (1987) and Vinson (1987) reviewed alternative methods for dealing with heterogeneous variances and discussed the potential difficulties of each method. Any adjustment procedure to account for heterogeneous variances and allow for an unbiased comparison of animals from various environments should be computationally feasible for large data sets.

Logarithmic transformation is often used to remove a relationship between mean and residual variance in fixed effects linear models. Log transformed production records are currently used in the Northeast AI Sire Comparison in an attempt to reduce the effect of heterogeneous error variance (Everett and Keown, 1984). Common variance and heritability estimates are used for all records and antilogs of sire solutions are obtained at the end of the analysis. Everett and Keown (1984) found that log transformation of data resulted in a reranking of bulls and produced greater differences among bulls than occurred with untransformed data. The researchers concluded that log transformation removes much of the relationship between mean and variance, and as a result, sire evaluations on log transformed data are superior to sire evaluations on untransformed data. The effect of log transformation on cow evaluation was not reported.

Log transformation of data is an appealing approach to the heterogeneous variance problem because it is computationally simple but log transformation alone may not effectively stabilize variances. De Veer and Van Vleck (1987) found that log transformation stabilized genetic variance but residual components of log yields decreased as the production level increased. Mirande and Van Vleck (1985) reported that on the log scale, residual standard deviations of milk yield were smallest with high production and largest with low production. They warned that use of a log transformation could lead to underevaluation of cows in high-production herds and overevaluation of cows in low-production herds. Results of simulation studies by Garrick and Van Vleck (1987) indicated that genetic progress in a dairy cattle population could be significantly decreased if variances for log transformed data were assumed to be homogeneous across all levels of production. Use of a single transformation to stabilize both genetic and residual variance may be a too simplistic approach.

Henderson (1984) has shown that his standard mixed model equations can be modified to account for nonstandard covariance structures such as heterogeneous genetic and residual variances. The appropriate form of the mixed model equations is determined by which variance components are heterogenous, i.e., genetic and/or residual. If only residual

variance is heterogeneous, a single trait model is applicable. If genetic components of variance are heterogeneous, a multiple trait model is used with genetic value in each herd treated as a different trait. This type of analysis would have high computing costs and would require the estimation of many covariances.

Based on previous studies (Danell, 1982; De Veer and Van Vleck, 1987; Hill et al., 1983), the appropriate model would be one in which additive genetic and residual variances and heritability are heterogeneous across environments but genetic correlation across environments is equal to one. This type of genotype-environment interaction is not one in which rank of animals change but instead results in greater absolute differences between evaluations of animals in high mean yield and variance herds in relation to low yield and variance herds. Henderson (1984) noted that under the assumption of unit genetic correlation between genetic expression in different environments, breeding values in each environment are linearly related and the mixed model equations are characterized by a singular genetic variance matrix. Gianola (1986) has shown, however, that if genetic correlation between merit in each environment is one, the multiple trait sire model can be reduced to a single trait model, which is computationally simpler. The first step of Gianola's method is to force the genetic variance to be equal for all records

by dividing each record by the appropriate genetic standard deviation. The resulting heterogenous residual variance matrix is then used in the mixed model equations.

Garrick and Van Vleck (1987) used sensitivity analyses to examine the increase in genetic gain resulting from modification of mixed model equations to account for heterogeneity of variances. For their example, actual sire and residual variances were assumed heterogeneous across three production levels. Sire variance increased at a faster rate than residual variance resulting in a doubling of heritability from the low to high variance environment. Evaluations were obtained for a model accounting for heterogeneous variances and for a model assuming homogeneity with estimates from the intermediate environment used for all environments. Results for a sire progeny test and bull dam selection indicated that little loss in genetic gain resulted from assuming variances were homogeneous. These findings support the work of Powell et al. (1983) who suggested that adjustments for larger variances and larger heritabilities in high mean yield herds would offset each other. The results of Garrick and Van Vleck (1987) are dependent, however, on the assumption of a two-fold increase in heritability from the low to high-production level which is greater than has been reported for most studies.

Two studies have incorporated adjustments for heterogeneous within-herd variances into mixed model genetic

evaluations for sires. Equivalent to the approach of Gianola (1986), Weller et al. (1985) used a double standardization to evaluate Israeli Holstein sires. In their approach, sire component of variance for each herd-year-parity group was computed as a function of the herd-year-parity mean. After dividing each record by the square root of the corresponding estimated sire variance, the inverse of the resulting residual variances were used to weight each record in the mixed model equations. The correlation between evaluations from a model assuming variances were homogeneous with the model accounting for heterogeneous variances was .99, but the average repeatability of the latter was slightly greater. The researchers suggested that it would be possible to extend this approach to other situations with unequal variance components. Weller et al. (1987) reported that the two-step standardization procedure is only slightly more difficult computationally than a sire evaluation model assuming homogeneity of variances.

The effect of heterogeneity of within-herd variances on the evaluation of Canadian Holstein sires was examined by Winkelman and Schaeffer (1988). Restricted maximum likelihood (REML) was used to estimate sire and residual variances for each herd. In contrast to most studies, variance estimates were not significantly correlated with herd production level. Data were randomly split into two subsets and sires were

evaluated within subsets by three models. In two models, a genetic correlation of one was assumed across herds and variances were assumed to be either homogeneous or heterogenous. Heterogeneous variances were also assumed in the third model but a genetic correlation of less than one was used in a multiple trait model described by Henderson (1984). The model in which heterogeneity of variances was ignored gave the greatest correlation between sire evaluations from the two data subsets. The researchers concluded that accounting for heterogeneous variances did not improve the accuracy of sire evaluation.

The results from field data suggest that modification of the mixed model equations to account for heterogeneous variances has little effect on the overall ranking of sires. The most important evaluations, however, are for sires and cows ranking in the top percentage of the population. These top sires are selected for extensive use throughout the cow population and top cows are candidates to produce young bulls for progeny testing. In comparison to sire evaluation, cow evaluation is more likely to be affected by heterogenous variances. The objective of this study was to compare the genetic evaluation of dairy sires and cows obtained from three alternative models. In two models, constant variances were assumed and the analysis used either untransformed or log transformed milk yields. In the third model, untransformed

yields were used and variances were assumed to be different for each of three production levels.

Materials and Methods

Description of data

The data previously described by Boldman (1988) was supplied by the United States Department of Agriculture Animal Improvement Programs Laboratory (USDA-AIPL) and consisted of 121,136 mature equivalent (ME), 2x, 305-d first and later lactation milk yields for 91,206 AI Holstein cows calving between 1979 and 1984 throughout the United States. Two seasons, November through April and May through October, were defined per year. Data consisted of daughters from 526 sires; 485 sires had first crop daughters in the data. Ninety-two common sires and maternal grandsires of the 526 sires with daughters were identified and included in the evaluation to account for relationships among the sires. Each record was assigned to one of three production levels based on the mean ME milk yield of all cows freshening in the same herd-year. Upper and lower limits used to define the three production levels were: low, 5897-7484 kg; medium, 7485-8618 kg; and high, 8619-10,206 kg. To avoid selection bias due to culling, cows without a first lactation record were eliminated; later lactations were included if they were assigned to the same production level as the first. Percentages of records assigned to low, medium, and high-production levels were 20.4,

47.8, and 31.9%, respectively. Other edits of the data are described by Boldman (1988).

Models

The assumed model for prediction of sire and cow breeding values in the three analyses included fixed herd-year-season (h) and sire genetic group (g) effects, and random animal (a) and permanent environment (p) within group, and residual (e) effects:

$$y_{ijmn} = h_i + g_j + a_m + p_m + e_{ijmn} \quad [1]$$

where y_{ijmn} is value of record n of animal m in sire genetic group j in herd-year-season i. Six sire genetic groups were defined as described by Boldman (1988). Sires were assigned to genetic groups by year of birth and cows were assigned to the same group as their sire. Additive genetic relationships from sires and maternal grandsires of sires and sires of cows were used but mates of sires were ignored. This approximate animal model which ignores mates of sires and female relationships is equivalent to the sire and nested cow model (see Appendix D for a numerical example) used by Boldman (1988) to estimate variance components for the same data, but the approximate animal model used to estimate breeding values allowed heterogeneous variances to be more easily incorporated. The model in [1] can be expressed in matrix notation as:

$$y = Xb + Za + Zp + e \qquad [2]$$

where y, b, a, p, and e are vectors of yields, fixed effects (herd-year-seasons and genetic groups), animal effects, permanent environment effects, and residual effects, respectively; and **X** and **Z** are design matrices for b and a and p, respectively. Under a no-selection model, E(y)=**X**b, E(a)= E(p)=E(e)=0, and

$$\text{Var}\begin{bmatrix} a \\ p \\ e \end{bmatrix} = \begin{bmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix}.$$

The mixed model equations (MME) for model [2] are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{P}^{-1} \end{bmatrix}\begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}y \\ \mathbf{Z}'\mathbf{R}^{-1}y \\ \mathbf{Z}'\mathbf{R}^{-1}y \end{bmatrix} \qquad [3]$$

If homogeneous genetic, permanent environmental, and residual components of variance are assumed for all records, **G** is the matrix of additive genetic relationships (denoted **A**) multiplied by additive genetic variance (σ_a^2), and **P** and **R** are diagonal matrices consisting of constant permanent environmental (σ_p^2) and residual (σ_e^2) variances, respectively. Multiplying the equations in [3] by the scalar residual variance σ_e^2 results in the equivalent equations:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z+A^{-1}\alpha_a & Z'Z \\ Z'X & Z'Z & Z'Z+I\alpha_p \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix} \quad [4]$$

where $\alpha_a=\hat{\sigma}^2_e/\hat{\sigma}^2_a$, the ratio of residual to additive genetic variance, and $\alpha_p=\hat{\sigma}^2_e/\hat{\sigma}^2_p$, the ratio of residual to permanent environmental variance.

Breeding values ($\hat{g}_j+\hat{a}_m$) for sires and cows were obtained from three analyses using the variance components estimated by REML from the same data (Boldman, 1988). In all analyses, the genetic correlation between breeding values in different production levels was assumed to be one. The analyses differed in the type of records, untransformed or natural log transformed, and whether heterogeneous variances were or were not considered.

In two of the analyses, heterogeneous variances were not considered; variance component estimates of milk yield at the medium-production level were used for all records in mixed model equations [4]. In the first of these two homogenous variance models, untransformed yields were used while log transformed yields were used in the second model. In the third analysis, untransformed yields were used and heterogenous variances at the three production levels were accounted for in the mixed model equations by the method of Gianola (1986). In this procedure, genetic variance is

standardized across production levels by dividing each record in [1] by the estimated genetic standard deviation corresponding to the production level of the herd:

$$y_{ijmn}/\sigma_{a_q} = (h_i + g_j + a_m + p_m + e_{ijmn})/\sigma_{a_q}$$

which is rewritten as:

$$y^*_{ijmn} = h^*_i + g^*_j + a^*_m + p^*_m + e^*_{ijmn} \qquad [5]$$

where σ_{a_q} is the genetic standard deviation for one of three production levels q. The variance-covariance matrices for the random elements of the model in [5] are:

$$\text{Var}\begin{bmatrix} a^* \\ p^* \\ e^* \end{bmatrix} = \begin{bmatrix} \mathbf{G}^* & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{P}^* & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R}^* \end{bmatrix}$$

where $\mathbf{G}^*=\mathbf{A}$, the matrix of additive genetic relationships; $\mathbf{P}^*$ is a diagonal matrix with elements $\sigma^2_{p_q}/\sigma^2_{a_q}$ for cows at production level q; and $\mathbf{R}^*$ is a diagonal matrix with elements $\sigma^2_{e_q}/\sigma^2_{a_q}$ for yields at production level q (see Appendix E for a numerical example). These variance-covariance matrices for the model [5] were used in mixed model equations:

$$\begin{bmatrix} \mathbf{X'R^{*-1}X} & \mathbf{X'R^{*-1}Z} & \mathbf{X'R^{*-1}Z} \\ \mathbf{Z'R^{*-1}X} & \mathbf{Z'R^{*-1}Z+G^{*-1}} & \mathbf{Z'R^{*-1}Z} \\ \mathbf{Z'R^{*-1}X} & \mathbf{Z'R^{*-1}Z} & \mathbf{Z'R^{*-1}Z+P^{*-1}} \end{bmatrix}\begin{bmatrix} \hat{b}^* \\ \hat{a}^* \\ \hat{p}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{*-1}}y^* \\ \mathbf{Z'R^{*-1}}y^* \\ \mathbf{Z'R^{*-1}}y^* \end{bmatrix} \qquad [6]$$

to estimate breeding values for cows and sires in analysis three. The rank of animal solutions from equations [6] are

the same as those obtained from a multiple trait model in which genetic merit at each production level is considered a different trait (Gianola, 1986). REML estimates of heritability, repeatability, and variance components (Boldman, 1988) used at each production level in the three analyses are in Table 1.

Breeding values from the adjusted variance model can be converted to breeding values for a particular production level by multiplying by the genetic standard deviation corresponding to that production level. For example, breeding values from the adjusted variance model are transformed to the medium production level by multiplying by 621 kg, the genetic standard deviation at the medium-production level.

Computational aspects

A traditional approach for solving large order mixed model equations is to explicitly form the equations and then iterate on the nonzero elements. This approach usually requires several complex programs and long computing time. Schaeffer and Kennedy (1986) presented an algorithm to obtain mixed model solutions without forming the system of equations. In their method, termed the indirect approach, solutions are obtained by Gauss-Seidel (G-S) iteration on the data. The basic ideal of the indirect approach is to read the data files and accumulate a diagonal element and an adjusted right-hand side (ARHS) for each effect in the model. Solutions for each

Table 1. Heritabilities, repeatabilities, and variance components by level of production used in the three mixed model analyses

Analysis (variances, scale)	Level of production		
	Low	Medium	High
1. unadjusted[a], linear kg			
heritability	0.22	0.22	0.22
repeatability	0.54	0.54	0.54
$\alpha_a=\hat{\sigma}^2_e/\hat{\sigma}^2_a$	2.08	2.08	2.08
$\alpha_p=\hat{\sigma}^2_e/\hat{\sigma}^2_p$	1.46	1.46	1.46
2. unadjusted[a], log kg			
heritability	0.21	0.21	0.21
repeatability	0.53	0.53	0.53
$\alpha_a=\hat{\sigma}^2_e/\hat{\sigma}^2_a$	2.28	2.28	2.28
$\alpha_p=\hat{\sigma}^2_e/\hat{\sigma}^2_p$	1.48	1.48	1.48
3. adjusted[b], linear kg			
heritability	0.18	0.22	0.24
repeatability	0.50	0.54	0.54
$\mathbf{P}^{*-1}=\mathbf{I}(\hat{\sigma}^2_{a_q}/\hat{\sigma}^2_{p_q})$	0.534	0.702	0.804
$\mathbf{R}^{*-1}=\mathbf{I}(\hat{\sigma}^2_{a_q}/\hat{\sigma}^2_{e_q})$	0.353	0.480	0.532
$1/\hat{\sigma}_{a_q}$	0.00197	0.00161	0.00140

[a]Variance estimates at medium production level used for all records.

[b]Heterogeneous variances across production levels adjusted by procedure of Gianola (1986).

effect are then obtained by dividing each ARHS by the corresponding diagonal element. Diagonal elements are the only part of the coefficient matrix explicitly formed and consist of number of records for an effect plus contributions from variances ratios if the effect is random. ARHS are yields adjusted by current solutions of all other effects in the record. In comparison to the traditional approach, the G-S indirect approach requires a single, less-complex program, but it also requires several sorted copies of the observation and relationship files. Misztal and Gianola (1987) presented an alternative indirect approach using a combination of G-S and Jacobi iteration. Their approach requires greater computer memory than the G-S indirect approach, but only one unsorted copy of the observation file and the relationship file are required and programming is further simplified.

A computer program written by Misztal (1987), based on the algorithm of Misztal and Gianola (1987), was modified to provide solutions in the three analyses. Gauss-Seidel iteration, which requires current solutions for each effect in computing following effects, was used for herd-year-season, group, and permanent environmental effects. Jacobi iteration, which does not require current solutions for preceding effects, was used for animal effects to simplify processing of relationship data. A relaxation factor of .8 (Misztal, 1987) was used for Jacobi iteration after the second round to speed

convergence of animal solutions. The relaxation factor was used to add a percentage of the difference between the previous two rounds to the solution for the current round.

To minimize input and output operations, the 205,201 solutions (22,525 herd-year-seasons, 6 sire genetic groups, 91,026 permanent environments, and 91,644 animals [618 sires and 91,026 cows]) and all data (121,136 observations and 91,644 pedigree records) were stored in approximately 7 Mb of memory. Convergence (C) was measured as sum of squared differences between solutions for successive rounds divided by sum of squared solutions for the current round (Schaeffer and Kennedy, 1986):

$$C^{(n+1)} = \frac{\Sigma(\text{solution}^{(n+1)} - \text{solution}^{(n)})^2}{\Sigma(\text{solution}^{(n+1)})^2} \qquad [7]$$

where n is the round of iteration and the summation is over all 205,201 solutions in the model. Schaeffer and Kennedy (1986) stated that a value of $1x10^{-10}$ or less indicates convergence while Misztal (1987) stated that over 100 rounds of iteration are required for an acceptable level of convergence. In each of the three analyses, 200 rounds of iteration were performed, at which point the convergence criterion was less than $1x10^{-10}$. Approximately 27 minutes total central processing unit time on a National Advanced Systems AS/9160 computer was required for iteration in each analysis.

Results and Discussion

Sires and cows were ranked on genetic evaluations ($\hat{g}_j+\hat{a}_m$) from the three models to determine if log transformation of yields or adjustment for heterogeneous variances resulted in important differences from a model assuming homogeneous variances. Table 2 presents a comparison of animal ranks from the model adjusted for heterogeneous variances (analysis 3) with animal ranks from the unadjusted linear (analysis 1) and unadjusted log yield (analysis 2) models. Rank correlations of evaluations from the three models were greater than .99 for both sires and cows, indicating a close relationship between the three evaluations. Figure 1 shows a high degree of similarity between ranks of the top 100 sires from the adjusted variance model and the unadjusted linear yield model. Weller et al. (1985) also reported a correlation of .99 for sire evaluations obtained from a model assuming equal variance components and a model adjusted for heterogeneous variances.

The evaluations of top sires and cows which are selected to enter AI studs and to be used as bull dams, respectively, are most important because they contribute most of the genetic gain in dairy populations. A comparison of the ranks of the top 5% of all sires and top 1% of all cows is also presented in Table 2. Ranks of top animals from the adjusted variance model were compared to ranks from the unadjusted variance models by three values, number in common and average and

Table 2. Comparison of animal ranks from the model adjusted for heterogeneous variances with animal ranks from the unadjusted linear and log yield models

	Models compared	
	Adjusted vs. unadjusted linear yield	Adjusted vs. unadjusted log yield
Sires		
Rank correlation	.99	.99
Top 5% (n=26)		
No. in common	23 (88%)	23 (88%)
Average rank difference	1.4	2.9
Maximum rank difference	7	10
Cows		
Rank correlation	.99	.99
Top 1% (n=910)		
No. in common	872 (96%)	781 (86%)
Average rank difference	48.6	185.3
Maximum rank difference	235	4227

maximum rank difference. For example, 23 of the top 26 sires (5%) from the adjusted variance model were also ranked in the top 5% by the unadjusted linear yield model. Average rank difference was calculated by subtracting from the rank of each top animal in the adjusted variance model the rank of the same animal in the unadjusted variance model. The absolute differences were summed over animals and the sum was then divided by the number of top animals to give the average rank difference. For example, for the top 26 sires, the average rank difference was calculated from:

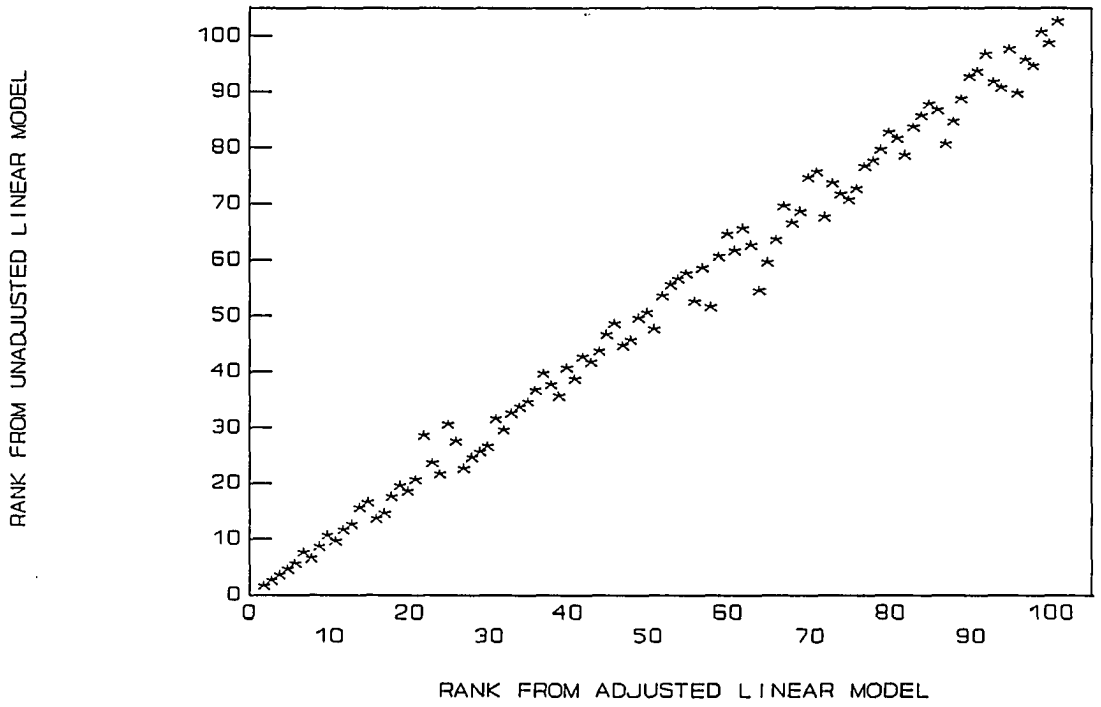

Figure 1. Comparison of sire ranks from the adjusted variance model and the unadjusted variance linear yield model

$$\text{Average rank difference} = \frac{\sum_{m=1}^{26} \left| \begin{array}{l}\text{rank of sire m} \\ \text{from adjusted} \\ \text{variance model}\end{array} - \begin{array}{l}\text{rank of sire m} \\ \text{from unadjusted} \\ \text{variance model}\end{array} \right|}{26} \quad [8]$$

where the summation was over the top 26 ranking sires from the adjusted variance model. The maximum rank difference was the largest change in rank for an individual animal. For example, the maximum cow rank change was 4227 for a cow whose rank was 860 in the variance adjusted model and 5087 in the unadjusted log yield model.

Ranks of top sires differed less than ranks of top cows across the three models. Heterogeneous variances should have

less effect on sire evaluation because daughters in low variance herds offset daughters in high variance herds. Only 23 of the top 26 sires from the adjusted variance model also ranked in the top 5% in the unadjusted linear and log yield models, indicating that the distribution of daughters was not random with respect to production level. Even though the number of sires in common with the adjusted variance model was the same for both the unadjusted linear and log yield models, the average rank difference for the log model was twice as large as the average difference for the linear model. Heterogeneity of variances could have a greater impact on the initial progeny test of a sire if a limited number of herds are represented.

As expected, cow evaluation was more sensitive to heterogeneity of variances. For cow evaluation, the unadjusted log yield model was inferior to the unadjusted linear yield model as indicated by a smaller number of elite cows in common with the adjusted variance model and also a larger average rank difference. Even though the unadjusted linear yield model and adjusted model identified most of the same elite cows, the average rank difference of 48.6 was large. In the model used, relationships through females were ignored. As a result, the evaluation of a cow was a function of the cow's own production records and the evaluation of her sire; the evaluation of her dam did not contribute. Vinson

(1987) noted that since dams and daughters usually produce in the same herd, biases resulting from heterogeneity of variances will be increased if the evaluation for a cow incorporates the evaluation of her dam. Therefore, effects of heterogeneity of variance on cow evaluation are potentially greater in an individual animal model where the evaluation of a cow incorporates the evaluation of her sire and dam.

The average and maximum rank difference between the adjusted model and the unadjusted linear yield model was smaller than the average rank difference between the adjusted model and the unadjusted log yield model for both sire and cow evaluation. Under the assumption that the model adjusted for heterogeneous variances is best, use of linear yields is recommended over use of log yields in a model in which homogeneous variances are assumed. These findings are in agreement with the simulation work of Garrick and Van Vleck (1987) which indicated that log transformation is a worse approach than simply ignoring heterogeneous variances.

The percentage of elite cows should increase in low yield herds and decrease in high yield herds after adjustment for heterogeneous variances. Table 3 presents for the three models the number and percentage of cows ranking in the top 1% by production level. For all models, the greatest percentage of elite cows was found at the high production level. For example, in the adjusted variance linear yield model, 0.57% of

the cows at the low production level classified as elite but 1.32% of the cows at the high production level classified as elite. Compared to the unadjusted variance linear yield model, the unadjusted log yield model and adjusted model increased the percentage of elite cows in low and medium mean production herds and decreased the percentage in high production herds. Changes in percentages of elite cows by production level were greater for the log yield model. For the log yield model, the percentage of elite cows was similar across the three levels of production, indicating an apparent excess and shortage of elite cows at the low and high level, respectively. As predicted by Mirande and Van Vleck (1985), use of log yield with an assumption of equal variances resulted in overevaluation of cows in low-production herds and underevaluation of cows in high-production herds, the opposite of the unadjusted linear yield model.

Table 3. Number and percentage of elite cows by production level for the three models

Level of production	Model of analysis		
	Unadjusted, linear yield	Unadjusted, log yield	Adjusted, linear yield
Low	102 (0.54%)[a]	164 (0.87%)	107 (0.57%)
Medium	386 (0.90%)	429 (1.00%)	419 (0.97%)
High	422 (1.45%)	317 (1.09%)	384 (1.32%)

[a]Elite cows expressed as a percentage of total number of cows at each production level in parentheses.

Based on the results of this study, the choice of a model for dairy cow and sire evaluation is between one in which heterogeneous variances are accounted for or one in which heterogeneity is ignored; log transformation of yields appears to be worse than no adjustment. Van Vleck (1988) pointed out that two issues exist when considering adjustments for heterogeneous variance in dairy evaluation models. First is the fairness issue, i.e., do all herds have an equal chance in producing bull dams? Adjusting for heterogeneous variances did increase the percentage of elite cows in low and medium-production herds (Table 3), but the largest percentage was found in high-production herds. This finding indicates that high-production herds result from better management and superior genotypes. The average breeding value for the adjusted variance model, scaled to the medium-production level by multiplying by $\hat{\sigma}_a$=621 kg, was -40.4, -12.1, and 30.1 kg for cows in low, medium, and high-production herds, respectively. Herds assigned on the basis of average phenotypic level also differ in average genetic merit.

The second question raised by Van Vleck (1988) is of greater importance and concerns the effect of adjustment for heterogeneous variances on genetic gain. According to Garrick and Van Vleck (1987), genetic progress in a dairy population might not be much affected by ignoring heterogeneous variances. On a population basis this would appear to be

nearly true because most of the top ranking cows and sires identified in the adjusted model were also identified in the unadjusted linear yield model, but the ranks of individual cows did change. Heterogeneous variance could result in the overevaluation of potential bull dams in high yield herds. If the subsequent progeny test is random across production levels, sires out of overevaluated cows should be identified as inferior and culled. In addition to the reduced selection intensity resulting from biases caused by heterogeneous variances, costs of sampling and housing bulls which do not enter active service are large.

A method to adjust for heterogeneous variances cannot be considered for use in a national sire and cow evaluation unless the method is computationally feasible. The approach of Gianola (1986) used in this study, in which a multiple trait analysis is reduced to a single trait model, requires few additional computations if accurate estimates of genetic, permanent environmental, and residual variances are available. If solutions are obtained by iteration on data, the variances associated with each record can be easily accounted for by weighting the contributions of each record when computing diagonals and ARHS of the effects in the model. The USDA national animal model evaluation for dairy cattle uses lactation length weights to determine contributions of records to diagonals and ARHS (Wiggans et al., 1988). Extension of

the weighting procedure to heterogeneous variances should be straightforward.

The main problem with the use of heterogeneous genetic and residual variance matrices in mixed model equations is that the variances must be estimated from limited data. In this study, only three levels of production were used. Additional levels would allow for greater precision in assigning herds, but fewer records would be available for estimation of variances at each level. Grouping herds on the basis of herd-year mean is simple, but within herd-year, mean and variance are only moderately correlated. Estimates of the correlation between within herd-year mean and standard deviation for milk yield have ranged from .24 (Everett et al., 1982) in data from the Northeast to .49 (Meinert, 1988) for national data. In a preliminary study of these data, a correlation of .34 was estimated from the full data set consisting of over 2 million records. Other herd characteristics could also be used to classify herds. A correlation of .15 between herd size and within-herd standard deviation was found for the full data set while a value of .28 was reported by Everett et al. (1982) for Northeast data.

An alternative approach would be to assign herds to groups on the basis of intraherd estimates of variance. Dong and Mao (1988) categorized herds into three groups by intraherd phenotypic standard deviation in an analysis to estimate

genetic and residual variances. A modification of this approach is to use estimates of genetic and residual variance from individual herds in mixed model equations. Winkelman and Schaeffer (1988) estimated sire and residual variances for individual herds by REML using across-herd estimates as priors. Estimated variances were not significantly correlated with herd mean or herd size. Estimating variance components for individual herds would be expensive in a large evaluation, and the estimates would have large sampling variances. Brotherstone and Hill (1986) reported that within-herd variance for the same herd was consistent over years so variances would need to be estimated only periodically. Henderson (1984) and Gianola (1986) have suggested a Bayesian approach to combine variance estimates for individual herds with pooled estimates from herds grouped by a characteristic such as production level. Gianola (1986) has warned, however, that the weights needed to combine variance estimates may be difficult to calculate in animal breeding data. For herds with a few records, the combined variance estimate would be largely determined by the population estimate. The most feasible approach to account for heterogeneous variances appears to be to assume constant variances for all records in a herd assigned to one of several subpopulations by production or phenotypic variance level.

In summary, the results indicate that sire evaluation is little affected by heterogeneous variances but cow evaluation is more sensitive to violations of the assumed homogeneous variance structure of an evaluation model. Log transformation of yields is a simple approach but is not recommended because cows in low yield herds are overevaluated.

The mixed model equations can be modified to include a two-step transformation to account for heterogeneous variances at several production or variance levels. This approach would be easy to implement in a national evaluation if reliable estimates of variance components were available.

References

Boldman, K. G. 1988. Heterogeneity of variances by herd production level and its effect on dairy cow and sire evaluation. Unpublished Ph.D. Dissertation. Iowa State University, Ames, IA.

Brotherstone, S. and W. G. Hill. 1986. Heterogeneity of variance amongst herds for milk production. Anim. Prod. 42:297.

Danell, B. 1982. Interaction between genotype and environment in sire evaluation for milk production. Acta Agric. Scand. 32:33.

De Veer, J. C. and L. D. Van Vleck. 1987. Genetic parameters for first lactation milk yields at three levels of herd production. J. Dairy Sci. 70:1434.

Dong, M. C. and I. L. Mao. 1988. Interaction between sire and level of intraherd production variance. J. Dairy Sci. 71(Suppl. 1):200. (Abstr.)

Everett, R. W. and J. F. Keown. 1984. Mixed model sire evaluation with dairy cattle-Experience and genetic gain. J. Anim. Sci. 59:529.

Everett, R. W., J. F. Keown, and J. F. Taylor. 1982. The problem of heterogeneous within herd error variances when identifying elite cows. J. Dairy Sci. 65(Suppl. 1):100. (Abstr.)

Garrick, D. J. and L. D. Van Vleck. 1987. Aspects of selection for performance in several environments with heterogeneous variances. J. Anim. Sci. 65:409.

Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogeneous. Theor. Appl. Genet. 72:671.

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada.

Hill, W. G., M. R. Edwards, M. K. A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition at different levels of variability of production. Anim. Prod. 36:59.

Meinert, T. R., R. E. Pearson, W. E. Vinson, and B. G. Cassel. 1988. Prediction of daughter's performance from dam's cow index adjusted for within-herd variance. J. Dairy Sci. 71:2220.

Mirande, S. L. and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.

Misztal, I. 1987. JAA: A Fortran program for solutions to a class of mixed models. User's guide. Dept. Anim. Sci., University of Illinois, Urbana.

Misztal, I. and D. Gianola. 1987. Indirect solution of mixed model equations. J. Dairy Sci. 70:716.

Powell, R. L., H. D. Norman, and B. T. Weinland. 1983. Cow evaluation at different milk yields of herds. J. Dairy Sci. 66:148.

Schaeffer, L. R. and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575.

Van Vleck, L. D. 1987. Contemporary groups for genetic evaluations. J. Dairy Sci. 70:2456.

Van Vleck, L. D. 1988. Alternatives for evaluations with heterogeneous genetic and environmental variances. J. Dairy Sci. 71(Suppl. 2):83. (Abstr.)

Vinson, W. E. 1987. Potential biases in genetic evaluations from differences in variation within herds. J. Dairy Sci. 70:2450.

Weller, J. I., M. Ron., and R. Bar-Anan. 1985. Accounting for environmentally dependent variance components in BLUP sire evaluations. J. Dairy Sci. 68(Suppl. 1):212. (Abstr.)

Weller, J. I., M. Ron., and R. Bar-Anan. 1987. Effects of persistency and production on the genetic parameters of milk and fat yield in Israeli-Holsteins. J. Dairy Sci. 79:672.

Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J. Dairy Sci. 71(Suppl. 2):54.

Winkelman, A. and L. R. Schaeffer. 1988. Heterogeneity of variances among herds and its effect on dairy sire evaluation. J. Dairy Sci. 71(Suppl. 2):84. (Abstr.)

RECOMMENDATIONS TO ACCOUNT FOR HETEROGENEITY OF VARIANCES IN A NATIONAL EVALUATION

The results from this study indicate that heterogeneity of variances may bias cow evaluation. In a national evaluation, several alternative approaches could be used to account for heterogeneous variances. Log transformation alone should not be used because environmental components of variance for transformed yields decreased as production level increased. On the other hand, a multiple trait approach is not needed because the results from this and other studies have indicated that the correlation between breeding values is close to one across environments. The proper approach to account for heterogeneity of variance depends on the ratio of residual to genetic variance, i.e., whether heritability is equal or not across herds.

Equal Heritability

In this situation, both additive genetic and residual variance differ from herd to herd but their ratio is assumed to be constant across herds. As a result, the total variance is different across herds but heritability is constant. A practical approach to account for heterogeneity would be to estimate phenotypic variance within each herd and then divide each observation by the corresponding estimated phenotypic standard deviation. As shown in Appendix E, if total variance is known, the solutions obtained from this approach are equivalent to those obtained from a two-step standardization

procedure, but unlike the latter approach, estimates of genetic and residual components of variance are not required. Estimation of phenotypic variance within single herds would be computationally feasible even in a large national evaluation but variance estimates based on few observations, e.g., estimates in small herds, would have large sampling variances. While this approach is straightforward, the assumption that heritability is equal across herds is probably unrealistic. Heritability of milk yield estimated in this study only increased from .18 in low-production herds to .24 in high-production herds, but most other studies have reported a larger increase in heritability. The approach of assuming equal heritability and dividing yields by estimated intra-herd phenotypic standard deviations when heritability is actually greater in high variance herds would result in the same problem as log transformation of yields, i.e., overevaluation of superior cows in low variance herds and underevaluation of superior cows in high yield herds.

Unequal Heritability

The assumption that both additive genetic variance and residual variance differ from herd to herd and heritability is greater in high variance herds is more realistic. In comparison to the equal heritability situation, the use of a single heritability value for all herds when this form of heterogeneity exits would probably result in a smaller

overevaluation of cows in high-variance herds because the low heritability value would offset the greater variances. In the presence of unequal heritability, a single standardization is not sufficient and estimates of both genetic and residual variance are required. Estimating both components for individual herds would be impractical and the estimated components would have large sampling variances. A feasible approach would be to assign each herd to one of several groups on the basis of within-herd standard deviation. Genetic and environmental components of variance would then be estimated within each group for use in the mixed model equations. Each record would be divided by the appropriate estimated genetic standard deviation to equalize genetic variance. Estimated permanent environmental and residual variance matrices, heterogeneous by variance level, would then be used in the mixed model equations. As indicated by the inability of log transformation of yields to stabilize variances, heterogeneity of variances is not simply the result of a relationship between mean and variance. Therefore, assignment of herds to variance groups should be done by estimated phenotypic standard deviation instead of by mean production.

SUMMARY

Results of a study of heterogeneity of production variances by herd production level and its effect on dairy cow and sire evaluation are presented. Methodology and results of the study are reported in two sections. Descriptions and numerical examples of several procedures used in the analysis are included in appendices.

In the first section, genetic, permanent environmental, and residual variances of untransformed and natural log transformed milk and fat yields were estimated from herd-year-seasons grouped into three levels of production. A univariate sire and nested cow model was used and estimation was by restricted maximum likelihood. In general, estimates of variance components for untransformed milk and fat yield increased with production level. Genetic variance increased at the greatest relative rate, resulting in the largest estimates of heritability at the high-production level. Log transformation did not stabilize estimates of variance components across the three levels. On the log scale, residual variance decreased as production level increased. Heritability estimates were not changed by log transformation of yields. A single transformation is unlikely to stabilize all variance components across all production levels. Correlations between sire values at different production levels were close to expected values, indicating that the

ranking of sires is not affected by the production level of their daughters.

In the second section, the estimated variance components were used in three mixed models to evaluate sires and cows for genetic merit of milk yield. The model used was equivalent to the sire and nested cow model used for variance component estimation in section one. In two models, variances estimated at the medium-production level were used for all records and the analysis used untransformed or log transformed milk yields. In the third model, heterogeneous variances of untransformed milk yield at the three production levels were used in the mixed model equations. Rank correlations of evaluations from the three models were close to unity for both sires and cows. Ranks of top sires were similar across models, but differences in ranks of top cows were large across the three models. Log transformation of milk yield appeared to result in an overevaluation of superior cows in low-production herds and underevaluation of superior cows in high herds and is a worse approach than simply ignoring heterogeneous variances. A model accounting for heterogeneous variances at several levels of production resulted in large changes in cow evaluation and is computationally feasible for large data sets if variance estimates are available.

REFERENCES

Aisbett, C. W. 1984. Association of herd means and variances is a function of edit for minimum lactation length. J. Dairy Sci. 67:702.

Blanchard, P. J., R. W. Everett, and S. R. Searle. 1983. Estimation of genetic trends and correlations for Jersey cattle. J. Dairy Sci. 66:1947.

Brotherstone, S. and W. G. Hill. 1986. Heterogeneity of variance amongst herds for milk production. Anim. Prod. 42:297.

Buttram, S. T. 1987. Genetics of racing performance in the American Quarter Horse. Unpublished Ph.D. Dissertation. Iowa State University, Ames, IA.

Calo, L. L., R. E. McDowell, L. D. Van Vleck, and P. D. Miller. 1973. Genetic aspects of beef production among Holstein-Friesians pedigree selected for milk production. J. Anim. Sci. 37:676.

Danell, B. 1982. Interaction between genotype and environment in sire evaluation for milk production. Acta Agric. Scand. 32:33.

De Veer, J. C. 1986. Genetic parameters for first lactation milk yields at three levels of production. Unpublished Ph.D. Dissertation. Cornell University, Ithaca, NY.

De Veer, J. C. and L. D. Van Vleck. 1987. Genetic parameters for first lactation milk yield at three levels of herd production. J. Dairy Sci. 70:1434.

Dong, M. C. and I. L. Mao. 1988. Interaction between sire and level of intraherd production variance. J. Dairy Sci. 71(Suppl. 1):200. (Abstr.)

Everett, R. W. and J. F. Keown. 1984. Mixed model sire evaluation with dairy cattle-Experience and genetic gain. J. Anim. Sci. 59:529.

Everett, R. W., J. F. Keown, and J. F. Taylor. 1982. The problem of heterogeneous within herd error variances when identifying elite cows. J. Dairy Sci. 65(Suppl. 1):100. (Abstr.)

Everett, R. W., R. L. Quaas, and E. J. Pollak. 1983. Multiple trait Northeast artificial insemination sire evaluation. Anim. Sci. Mimeo Series No. 74. Cornell University, Ithaca, NY.

Falconer, D. S. 1952. The problem of environment and selection. Am. Nat. 86:293.

Falconer, D. S. 1981. Introduction to quantitative genetics. Longman, Inc., New York, NY.

Garrick, D. J. and L. D. Van Vleck. 1987. Aspects of selection for performance in several environments with heterogeneous variances. J. Anim. Sci. 65:409.

Gianola, D. 1986. On selection criteria and estimation of parameters when the variance is heterogeneous. Theor. Appl. Genet. 72:671.

Harville, D. A. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Statist. Assoc. 72:320.

Henderson, C. R. 1975a. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423

Henderson, C. R. 1975b. Inverse of a matrix of relationships due to sires and maternal grandsires. J. Dairy Sci. 58:1917.

Henderson, C. R. 1984a. Applications of linear models in animal breeding. University of Guelph, Guelph, Canada.

Henderson, C. R. 1984b. Recent developments in variance and covariance estimation. J. Anim. Sci. 63:2082.

Henderson, C. R. 1985. Equivalent linear models to reduce computations. J. Dairy Sci. 68:2267.

Hickman, C. G., A. J. Lee, and K. Gravier. 1969. Genotype x season x method interaction in evaluating dairy sires from progeny records. Can. J. Anim. Sci. 49:151.

Hill, W. G., M. R. Edwards, M. K. A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition at different levels of variability of production. Anim. Prod. 36:59.

Hudson, G. S., R. L. Quaas, and L. D. Van Vleck. 1982. Computer algorithm for the recursive method of calculating large numerator relationship matrices. J. Dairy Sci. 26:2018.

Laird, N. M. and J. H. Ware. 1982. Random effects models for longitudinal data. Biometrics 38:963.

Lawlor, T. J. 1984. Estimation of genetic and phenotypic parameters of milk, fat and protein yields of Holstein cattle under selection. Unpublished Ph.D. Dissertation. Cornell University, Ithaca, NY.

Meinert, T. R., R. E. Pearson, W. E. Vinson, and B. G. Cassel. 1988. Prediction of daughter's performance from dam's cow index adjusted for within-herd variance. J. Dairy Sci. 71:2220.

Meyer, K. 1987. Restricted maximum likelihood to estimate variance components for mixed models with two random factors. Génét. Sél. Evol. 19(1):49.

Mirande, S. L. and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.

Misztal, I. 1987. JAA: A Fortran program for solutions to a class of mixed models. User's guide. Dept. Anim. Sci., University of Illinois, Urbana.

Misztal, I. and D. Gianola. 1987. Indirect solution of mixed model equations. J. Dairy Sci. 70:716.

Ouweltjes, W., L. R. Schaeffer, and B. W. Kennedy. 1988. Sensitivity of methods of variance component estimation to culling type of selection. J. Dairy Sci. 71:773.

Powell, R. L., H. D. Norman, and B. T. Weinland. 1983. Cow evaluation at different milk yields of herds. J. Dairy Sci. 66:148.

Robertson, A. 1977. The effect of selection on the estimation of genetic parameters. Z. Tierz. Zuchtungsbiol. 94:131.

Schaeffer, L. R. and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575.

Searle, S. R. 1982. Matrix algebra useful for statistics. John Wiley and Sons, Inc., New York, NY.

Sorenson, D. A. and B. W. Kennedy. 1984. Estimation of genetic variances from unselected and selected populations. J. Anim. Sci. 59:1213.

Thompson, R. and K. Meyer. 1986. Estimation of variance components: What is missing in the EM algorithm? J. Statist. Comput. Simul. 24:215.

Ufford, G. R., C. R. Henderson, and L. D. Van Vleck. 1978. Deviation of computing algorithms for sire evaluation using all lactation records and natural service sires. Anim. Sci. Mimeo Series No. 39. Cornell University, Ithaca, NY.

Van Vleck, L. D. 1977. Theoretical and actual genetic progress in dairy cattle. In E. Pollak, O. Kempthorne, and T. B. Bailey, Jr. (eds.). Proc. Int. Conf. Quant. Genet. Iowa State University Press, Ames, IA.

Van Vleck, L. D. 1987. Contemporary groups for genetic evaluations. J. Dairy Sci. 70:2456.

Van Vleck, L. D. 1988. Alternatives for evaluations with heterogeneous genetic and environmental variances. J. Dairy Sci. 71(Suppl. 2):83. (Abstr.)

VanRaden, P. M. and A. E. Freeman. 1987. Rates of convergence of REML algorithms. Iowa Agric. and Home Econ. Exp. Stn., Journal Paper No. J-12584.

VAST-E User's Guide. 1988. Edition 1.4. Pacific-Sierra Corp., Los Angeles, CA.

Vinson, W. E. 1987. Potential biases in genetic evaluations from differences in variation within herds. J. Dairy Sci. 70:2450.

Weller, J. I., M. Ron., and R. Bar-Anan. 1985. Accounting for environmentally dependent variance components in BLUP sire evaluations. J. Dairy Sci. 68(Suppl. 1):212. (Abstr.)

Weller, J. I., M. Ron., and R. Bar-Anan. 1987. Effects of persistency and production on the genetic parameters of milk and fat yield in Israeli-Holsteins. J. Dairy Sci. 79:672.

Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988a. Animal model evaluation of Ayrshire milk yield with all lactations, herd-sire interaction, and groups based on unknown parents. J. Dairy Sci. 71:1319.

Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988b. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J. Dairy Sci. 71(Suppl. 2):54.

Winkelman, A. and L. R. Schaeffer. 1988. Heterogeneity of variances among herds and its effect on dairy sire evaluation. J. Dairy Sci. 71(Suppl. 2):84. (Abstr.)

ACKNOWLEDGEMENTS

Appreciation is extended to USDA-AIPL for the data used in the study. Financial support for this study was provided by Eastern AI Cooperative and the National Association of Animal Breeders.

This thesis would not have been possible without the assistance of many people. Karin Meyer and Ignacy Misztal generously provided computer programs which were modified for variance component estimation and prediction of breeding values. I extend my appreciation to Drs. Aitchison, Berger, Harville, and Rothschild for serving on my committee and for all that I have learned in their classes. Dr. Freeman was always available when I needed assistance. He stimulated my interest in research and provided an excellent environment in which to work and learn. I feel fortunate to have had the opportunity to share an office with Paul VanRaden; his knowledge of animal breeding contributed much to my education. The computing assistance provided by Dr. Hoekstra was invaluable and Mohamed Sadek spent many hours on the initial edits of the data.

The friendship I found during my stay at Iowa State will never be forgotten. The staff, faculty, and fellow grad students in the department were a joy to work and play with. Special thanks to Dave Kelley. He was always available when I needed advice or just someone to talk to.

My parents stressed the importance of education and supported me throughout my studies. Most of all, I want to express my love and appreciation to my wife, Amy. She typed my thesis, but more importantly, she was always supportive and understanding. I could not have completed my thesis without her help and encouragement.

APPENDIX A. A REML ALGORITHM TO ESTIMATE VARIANCE COMPONENTS FOR A SIRE AND NESTED COW MODEL

Restricted Maximum Likelihood (REML) has become accepted as the preferred method to estimate variance components for animal breeding data. Henderson (1984a) described an expectation maximization (EM) type REML algorithm based on his mixed model equations (MME). Though the EM algorithm is often slow to converge, it yields non-negative estimates. Most univariate applications of REML have been restricted to models which include only a single random factor besides the error. Recently, Meyer (1987) presented an EM like algorithm for a univariate model with two random factors such as sires and cows nested within sires which is common in dairy cattle data. In this procedure, the trace of the inverse corresponding to cows is determined indirectly so the method is feasible for relatively large data sets.

Let the model of analysis include fixed herd-year-season (h) and sire genetic group (g) effects in addition to random sire (s), cow (c), and residual (e) effects:

$$y_{ijklm} = h_i + g_j + s_{jk} + c_{jkl} + e_{ijklm} \qquad [1]$$

with $E(y_{ijklm})=h_i+g_j$, $E(s_{jk})=E(c_{jkl})=E(e_{ijklm})=0$, and

$$\text{Var}\begin{bmatrix} s \\ c \\ e \end{bmatrix} = \begin{bmatrix} A\sigma_s^2 & 0 & 0 \\ 0 & I\sigma_c^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix}$$

where y_{ijklm} is value of record m of cow l of sire k in genetic group j in herd-year-season i, **A** is a matrix of additive relationships among sires, and σ_s^2, σ_c^2, and σ_e^2 are variances of s, c, and e effects, respectively. Cows are assumed unrelated to other cows and sires to simplify computations. The MME for model [1] are:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Q} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Q'X} & \mathbf{Q'Q} & \mathbf{Q'Z} & \mathbf{Q'W} \\ \mathbf{Z'X} & \mathbf{Z'Q} & \mathbf{Z'Z}+\mathbf{A}^{-1}\alpha_s & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Q} & \mathbf{W'Z} & \mathbf{W'W}+\mathbf{I}\alpha_c \end{bmatrix} \begin{bmatrix} \hat{h} \\ \hat{g} \\ \hat{s} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Q'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{bmatrix} \qquad [2]$$

where **X**, **Q**, **Z**, and **W** are design matrices for h, g, s, and c effects, respectively, $\alpha_s=\hat{\sigma}_e^2/\hat{\sigma}_s^2$, and $\alpha_c=\hat{\sigma}_e^2/\hat{\sigma}_c^2$. Total variance, σ_y^2, is the sum of sire, cow, and residual variances, i.e., $\sigma_y^2=\sigma_s^2+\sigma_c^2+\sigma_e^2$. The variance ratios required in the MME can be expressed as a function of heritability, $h^2=4\sigma_s^2/\sigma_y^2$, and repeatability, $r=(\sigma_s^2+\sigma_c^2)/\sigma_y^2$. Residual variance can be expressed as a function of total variance and repeatability:

$$\sigma_e^2=\sigma_y^2-(\sigma_s^2+\sigma_c^2)$$
$$\sigma_e^2=\sigma_y^2-r\sigma_y^2$$
$$\sigma_e^2=(1-r)\sigma_y^2 \ .$$

Similarly, sire variance can be expressed as a function of total variance and heritability, $\sigma_s^2=.25h^2\sigma_y^2$, and cow variance can be expressed as a function of total variance, heritability, and repeatability:

$$\sigma_c^2=r\sigma_y^2-\sigma_s^2$$
$$\sigma_c^2=r\sigma_y^2-.25h^2\sigma_y^2$$
$$\sigma_c^2=(r-.25h^2)\sigma_y^2 \ .$$

Therefore, the variance ratios can be expressed as $\alpha_s=\hat{\sigma}_e^2/\hat{\sigma}_s^2=(1-r)/(.25h^2)$ and $\alpha_c=\hat{\sigma}_e^2/\hat{\sigma}_c^2=(1-r)/(r-.25h^2)$.

Estimates of variance components can be obtained through an iterative scheme using the MME. Three estimating equations similar to the EM algorithm are (Meyer, 1987):

$$\hat{\sigma}_s^2 = \hat{s}'\mathbf{A}^{-1}\hat{s} \ /[ns - \alpha_s trace(\mathbf{A}^{-1}\mathbf{C}_{ss})] \ , \qquad [3]$$

$$\hat{\sigma}_c^2 = \hat{c}'\hat{c} \ /[nc - \alpha_c trace(\mathbf{C}_{cc})] \ , \text{ and} \qquad [4]$$

$$\hat{\sigma}_e^2 = \hat{e}'\hat{e} \ /[ndfe + \alpha_s trace(\mathbf{A}^{-1}\mathbf{C}_{ss}) + \alpha_c trace(\mathbf{C}_{cc})] \qquad [5]$$

where n, ns, and nc are the number of observations, sires, and cows, respectively; ndfe=n-ns-nc-rank(**X**:**Q**) denotes the degrees of freedom for error; $\mathbf{C}_{ss}$ and $\mathbf{C}_{cc}$ correspond to the sire and cow sections of the inverse of the left-hand sides of the MME; and $\hat{e}'\hat{e}=y'(y-\mathbf{X}\hat{h}-\mathbf{Q}\hat{g}-\mathbf{Z}\hat{s}-\mathbf{W}\hat{c})-\alpha_s\hat{s}'\mathbf{A}^{-1}\hat{s}-\alpha_c\hat{c}'\hat{c}$. Therefore, in order to estimate variance components, the EM algorithm requires solutions to the MME, the trace of $\mathbf{C}_{cc}$, and the trace of the product of $\mathbf{C}_{ss}$ and the inverse of the relationship matrix.

An efficient strategy is to first absorb cows and then absorb herd-year-seasons into sires and groups. If cows do

not change herds, repeated records for a cow are nested within herds and the inverse required to absorb herd-year-seasons can be calculated at the end of each herd. Because sires are nested within genetic groups, equations for genetic groups can be easily formed from the sire equations after the other effects have been absorbed. The MME for herd-year-seasons, sires, and cows (omitting groups) are:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z+A^{-1}\alpha_s & Z'W \\ W'X & W'Z & W'W+I\alpha_c \end{bmatrix} \begin{bmatrix} \hat{h} \\ \hat{s} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix} . \qquad [6]$$

After absorbing the cow effects, the equations in [6] can be written as:

$$\begin{bmatrix} X'BX & X'BZ \\ Z'BX & Z'BZ+A^{-1}\alpha_s \end{bmatrix} \begin{bmatrix} \hat{h} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X'By \\ Z'By \end{bmatrix} \qquad [7]$$

$$\text{with } B = I - W(W'W+I\alpha_c)^{-1}W' . \qquad [8]$$

Absorbing herd-year-season equations gives the MME (without groups) for sires as:

$$\begin{bmatrix} Z'PZ+A^{-1}\alpha_s \end{bmatrix} \begin{bmatrix} \hat{s} \end{bmatrix} = \begin{bmatrix} Z'Py \end{bmatrix} \qquad [9]$$

$$\text{with } P = B - BX(X'BX)^{-1}X'B . \qquad [10]$$

Group equations are formed by summing over the sire equations for each group. The absorbed MME for groups and sires are:

$$\begin{bmatrix} Q'PQ & Q'PZ \\ Z'PQ & Z'PZ+A^{-1}\alpha_s \end{bmatrix}\begin{bmatrix} \hat{g} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} Q'Py \\ Z'Py \end{bmatrix} . \qquad [11]$$

Therefore, after absorption a generalized inverse of order equal to the number of groups and sires is required,

$$\begin{bmatrix} Q'PQ & Q'PZ \\ Z'PQ & Z'PZ+A^{-1}\alpha_s \end{bmatrix}^{-} = \begin{bmatrix} C_{gg} & C_{gs} \\ C_{sg} & C_{ss} \end{bmatrix} \qquad [12]$$

to obtain group and sire solutions:

$$\begin{bmatrix} \hat{g} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} C_{gg} & C_{gs} \\ C_{sg} & C_{ss} \end{bmatrix}\begin{bmatrix} Q'Py \\ Z'Py \end{bmatrix} . \qquad [13]$$

At this point the group and sire solutions and the inverse of the coefficient matrix corresponding to sires, C_{ss}, have been calculated. In addition, the REML algorithm requires solutions for herd-year-seasons and cows. These solutions can be obtained for each herd by backsolving using matrices derived during absorption:

$$\hat{h} = (X'BX)^{-1}[X'By - X'BZ(\hat{s}+\hat{g})] \qquad [14]$$

$$\hat{c} = (W'W+I\alpha_c)^{-1}[W'y - W'X\hat{h} - W'Z(\hat{s}+\hat{g})] . \qquad [15]$$

The trace of C_{cc}, the inverse of the coefficient matrix corresponding to cows, can be derived by the use of partitioned matrix results, utilizing inverses and matrix products from the absorption steps.

Let:

$$H_c = (W'W+I\alpha_c)^{-1} , \quad [16]$$

$$H_h = (X'BX)^{-1} , \quad [17]$$

$$L_{hc} = X'WH_c , \quad [18]$$

$$L_{g:sc} = (Q:Z)'WH_c , \quad [19]$$

$$L_{g:sh} = (Q:Z)'BXH_h , \text{ and} \quad [20]$$

$$T = [L_{g:sc} - (L_{g:sh}L_{hc})][L_{g:sc} - (L_{g:sh}L_{hc})]' . \quad [21]$$

The trace of the inverse corresponding to cows is then:

$$tr(C_{cc}) = tr(H_c) + tr(H_hL_{hc}L'_{hc}) + tr(C_{ss}T) . \quad [22]$$

To illustrate the computations required to obtain the solutions and traces of the inverse of the MME used in the EM algorithm, consider the following data consisting of repeated records by ten cows of four sires in two herds:

Herd	Sire	Cow	No. Records	Records by Year-Season 1	2	3
1	1	1	3	1990	2081	2084
1	2	2	3	1255	1252	1214
1	2	3	3	2149	1455	1259
1	4	4	2	-	1689	1548
1	4	5	1	-	-	1704
2	1	6	1	-	1707	-
2	1	7	3	1347	1081	1478
2	3	8	1	-	1448	-
2	3	9	1	-	-	1877
2	3	10	3	1188	1355	1135

Sires 1 and 2 are in genetic group one and sires 3 and 4 are in group two. Let the model in [1] be used to analyze the data. Sires are unrelated, i.e., **A=I**, α_s=7.5, and α_c=1 from h^2=.25 and r=.53125. The full MME from [2] are:

$$
\begin{bmatrix}
3 \\
0 & 4 \\
0 & 0 & 5 \\
0 & 0 & 0 & 2 \\
0 & 0 & 0 & 0 & 4 \\
0 & 0 & 0 & 0 & 0 & 3 & & & \text{symmetric} \\
3 & 3 & 3 & 1 & 2 & 1 & 13 \\
0 & 1 & 2 & 1 & 2 & 2 & 0 & 8 \\
1 & 1 & 1 & 1 & 2 & 1 & 7 & 0 & 7+7.5 \\
2 & 2 & 2 & 0 & 0 & 0 & 6 & 0 & 0 & 6+7.5 \\
0 & 0 & 0 & 1 & 2 & 2 & 0 & 5 & 0 & 0 & 5+7.5 \\
0 & 1 & 2 & 0 & 0 & 0 & 0 & 3 & 0 & 0 & 0 & 3+7.5 \\
1 & 1 & 1 & 0 & 0 & 0 & 3 & 0 & 3 & 0 & 0 & 0 & \text{diag}(3+1 \\
1 & 1 & 1 & 0 & 0 & 0 & 3 & 0 & 0 & 3 & 0 & 0 & 3+1 \\
1 & 1 & 1 & 0 & 0 & 0 & 3 & 0 & 0 & 3 & 0 & 0 & 3+1 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 2 & 2+1 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1+1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1+1 \\
0 & 0 & 0 & 1 & 1 & 1 & 3 & 0 & 3 & 0 & 0 & 0 & 3+1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1+1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1+1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 3 & 0 & 0 & 3 & 0 & 3+1)
\end{bmatrix}
\begin{bmatrix}
\hat{h}_1 \\ \hat{h}_2 \\ \hat{h}_3 \\ \hat{h}_4 \\ \hat{h}_5 \\ \hat{h}_6 \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \\ \hat{s}_4 \\ \hat{c}_1 \\ \hat{c}_2 \\ \hat{c}_3 \\ \hat{c}_4 \\ \hat{c}_5 \\ \hat{c}_6 \\ \hat{c}_7 \\ \hat{c}_8 \\ \hat{c}_9 \\ \hat{c}_{10}
\end{bmatrix}
=
\begin{bmatrix}
5394 \\ 6477 \\ 7809 \\ 2535 \\ 5591 \\ 4490 \\ 20352 \\ 11944 \\ 11768 \\ 8584 \\ 7003 \\ 4941 \\ 6155 \\ 3721 \\ 4863 \\ 3237 \\ 1704 \\ 1707 \\ 3906 \\ 1448 \\ 1877 \\ 3678
\end{bmatrix}.
$$

A generalized inverse of the left-hand side is required to give a set of solutions:

$\hat{h}_i$ = (673.2 477.6 396.4 222.9 242.8 343.5) ,

$\hat{g}_j$ = (1146.1 1210.3) ,

$\hat{s}_{jk}$ = (38.5 -38.5 -3.4 3.4) ,

$\hat{c}_{jkl}$ = (263.5 -287.2 -1.7 -21.5 47.0 139.8 -114.3
-0.9 163.3 -188.0) ,

and traces of the sire and cow portions of the inverse trace(C_{ss})=0.503 and trace(C_{cc})=5.906.

This approach in which the left-hand side of the full mixed model equations is inverted to obtain solutions and traces required in the EM algorithm is not computationally feasible for models with many factors. The alternative approach in which solutions are obtained by absorption and back solution and the trace of C_{cc} is determined indirectly is computationally feasible for many large models.

The first step in the alternative method is to absorb each cow into the herd-year-season and sire equations. If cows are assumed unrelated, this can be done without inversion by accumulating six quantities for each cow (Ufford et al., 1978). If the data are sorted by herds, herd-year-seasons can be absorbed into sires at the end of each herd. The absorbed MME for sires and groups are from [11]:

$$\begin{bmatrix} 1.437 & & & & & \\ -1.437 & 1.437 & & \text{symmetric} & & \\ 0.954 & -0.954 & 1.293+7.5 & & & \\ 0.483 & -0.483 & -0.339 & 0.822+7.5 & & \\ -0.712 & 0.712 & -0.712 & 0 & 0.712+7.5 & \\ -0.725 & 0.725 & -0.242 & -0.483 & 0 & 0.725+7.5 \end{bmatrix} \begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{s}_3 \\ \hat{s}_4 \end{bmatrix} = \begin{bmatrix} -74.16 \\ 74.16 \\ 292.22 \\ -366.38 \\ -9.66 \\ 83.83 \end{bmatrix}.$$

A generalized inverse of the absorbed left-hand side gives the same sire solutions and the same difference between group solutions (64.2) as from the full MME:

$$\hat{g}_j = (\ -32.1 \quad 32.1 \ ) \text{ and}$$

$$\hat{s}_{jk} = (\ 38.5 \quad -38.5 \quad -3.4 \quad 3.4 \ )$$

and the trace of C_{ss}=0.503.

Herd-year-season and cow solutions are obtained by backsolving using [14] and [15]:

$\hat{h}_i$ = (1851.4 1655.8 1574.6 1401.0 1421.0 1521.7) and

$\hat{c}_{jkl}$ = (263.5 -287.2 -1.7 -21.5 47.0 139.8 -114.3 -0.9 163.3 -188.0).

The backsolved herd-year-season solutions are not the same as from the full MME but the estimable functions are the same for both sets of solutions, e.g., $\hat{h}_1-\hat{h}_2$=195.6.

The trace of C_{cc} required in the EM algorithm is determined indirectly using inverses and matrix products from the absorption steps:

$$H_c = (W'W+I\alpha_c)^{-1} \qquad [16]$$

$$= \text{diag} \ (.25 \ .25 \ .25 \ .33 \ .50 \ .50 \ .25 \ .50 \ .50 \ .25) \ ,$$

$$H_h = (X'BX)^{-1} \qquad [17]$$

$$= \begin{bmatrix} .586 & .225 & .2 & & & \\ .225 & .475 & .2 & & \mathbf{0} & \\ .2 & .2 & .4 & & & \\ & & & .826 & .217 & .261 \\ & \mathbf{0} & & .217 & .478 & .174 \\ & & & .261 & .174 & .609 \end{bmatrix},$$

$$L_{hc} = X'WH_c \qquad [18]$$

$$= \begin{bmatrix} .25 & .25 & .25 & 0 & 0 & & & & & \\ .25 & .25 & .25 & .33 & 0 & & \mathbf{0} & & & \\ .25 & .25 & .25 & .33 & .5 & & & & & \\ & & & & & 0 & .25 & 0 & 0 & .25 \\ & & \mathbf{0} & & & .5 & .25 & .5 & 0 & .25 \\ & & & & & 0 & .25 & 0 & .5 & .25 \end{bmatrix},$$

$$L_{g:sc} = (Q:Z)'WH_c \qquad [19]$$

$$= \begin{bmatrix} .75 & .75 & .75 & 0 & 0 & .5 & .75 & 0 & 0 & 0 \\ 0 & 0 & 0 & .67 & .5 & 0 & 0 & .5 & .5 & .75 \\ .75 & 0 & 0 & 0 & 0 & .5 & .75 & 0 & 0 & 0 \\ 0 & .75 & .75 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .5 & .5 & .75 \\ 0 & 0 & 0 & .67 & .5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$L_{g:sh} = (Q:Z)'BXH_h \qquad [20]$$

$$= \begin{bmatrix} .758 & .675 & .6 & .435 & .457 & .348 \\ .242 & .325 & .4 & .565 & .543 & .652 \\ .253 & .225 & .2 & .435 & .457 & .348 \\ .506 & .450 & .4 & 0 & 0 & 0 \\ 0 & 0 & 0 & .565 & .543 & .652 \\ .242 & .325 & .4 & 0 & 0 & 0 \end{bmatrix}, \text{ and}$$

$$T = [L_{g:sc} - (L_{g:sh}L_{hc})][L_{g:sc} - (L_{g:sh}L_{hc})]' \qquad [21]$$

$$= \begin{bmatrix} .892 & & & & & \\ -.892 & .892 & & & & \\ .595 & -.595 & .870 & \text{symmetric} & & \\ .297 & -.297 & -.276 & .573 & & \\ -.446 & .446 & -.446 & 0 & -.446 & \\ -.446 & .446 & -.149 & -.297 & 0 & .446 \end{bmatrix}.$$

The trace of the inverse corresponding to cows is:

$$tr(C_{cc}) = tr(H_c) + tr(H_h L_{hc} L'_{hc}) + tr(C_{ss}T) \qquad [22]$$

$$= 3.583 + 1.544 + .779$$

$$= 5.906$$

which is the same as from the inverse of the full MME. The solutions to and traces of the MME obtained by the alternative approach are used in equations [3], [4], and [5] to estimate sire, cow, and residual variances respectively by the EM

algorithm.

Under the assumption that cows are unrelated through females and nested within herds, cows can be absorbed one at a time. The strategy requires the inverse of a matrix of order equal to the number of herd-year-seasons for each herd, but this would not be large for most data sets. The limiting step would most likely be the inversion of the absorbed sire equations to obtain C_{ss}.

APPENDIX B. ALTERNATIVE SIRE QUADRATICS USING THE EXPECTATION MAXIMIZATION ALGORITHM AND RELATIONSHIPS AMONG SIRES

A single trait sire model commonly used in dairy cattle breeding to estimate variance components and evaluate sires is:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{s} + \mathbf{e} \qquad [1]$$

with $E[\mathbf{y}]=\mathbf{X}\beta$, $E[\mathbf{s}]=E[\mathbf{e}]=\mathbf{0}$, and

$$\text{Var}\begin{bmatrix} \mathbf{s} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_s^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

where y is the data vector, β is the vector of fixed effects, s is the vector of random sire effects, **X** and **Z** are incidence matrices associated with vectors β and s, respectively; e is the vector of random residuals, **A** is the numerator relationship matrix, and σ_s^2 and σ_e^2 are variances of s and e effects, respectively.

The inverse of the relationship matrix, $\mathbf{A}^{-1}$, is often used in variance component estimation procedures. REML incorporates the inverse of the relationship matrix for computing quadratics and for expectations of quadratics. Use of relationships is especially important when selection of data has been practiced. Henderson (1975a) has shown that BLUP under a selection model is the same as BLUP ignoring selection provided certain conditions are met. One condition is that the variances and covariances of the population prior

to selection be known. Selection tends to decrease additive genetic variance so use of selected data will yield a biased estimate of the base population additive genetic variance. Allowing for relationships among sires should reduce the bias, however (Sorenson and Kennedy, 1984). Lawlor (1984) used REML and a sire model and reported that estimates of heritability increased from .17 to .19 for milk yield and from .22 to .26 for fat yield when relationships among sires were used. C. R. Henderson (personal communication, Dept. of Animal Sci., Univ. of Illinois, Urbana, 1988) has proven that ignoring relationships that exist will result in a reduction in REML estimates of genetic variance. This result is a compelling argument for including relationships when estimating variance components by REML, especially in data arising from selection.

Henderson's (1975b) procedure for generating $\mathbf{A}^{-1}$ expands the sire vector, s, to include sires and maternal grandsires of the sires in the analysis. If the sire vector is partitioned into sires without and sires with daughters, the model in [1] can be rewritten as:

$$y = \mathbf{X}\beta + \begin{bmatrix} 0 & \mathbf{Z} \end{bmatrix} \begin{bmatrix} s_0 \\ s_1 \end{bmatrix} + e \qquad [2]$$

where s_0 is a vector of sires without daughters and s_1 is a vector of sires with daughters and the variance of the

partitioned sire vector is:

$$\text{Var}\begin{bmatrix} s_0 \\ s_1 \end{bmatrix} = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix} \sigma_s^2 .$$

The mixed model equations (MME) for model [2] are:

$$\begin{bmatrix} X'X & 0 & X'Z \\ 0 & A^{00}\alpha & A^{01}\alpha \\ Z'X & A^{10}\alpha & Z'Z+A^{11}\alpha \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{s}_0 \\ \hat{s}_1 \end{bmatrix} = \begin{bmatrix} X'y \\ 0 \\ Z'y \end{bmatrix} \qquad [3]$$

where

$$\begin{bmatrix} A^{00} & A^{01} \\ A^{10} & A^{11} \end{bmatrix} = \begin{bmatrix} A_{00} & A_{01} \\ A_{10} & A_{11} \end{bmatrix}^{-1} = A^{-1}$$

and $\alpha=\sigma_e^2/\sigma_s^2$.

A quadratic for the sire effects using the expectation maximization (EM) algorithm (Henderson, 1984a) is:

$$\begin{bmatrix} \hat{s}_0' & \hat{s}_1' \end{bmatrix} \begin{bmatrix} A^{00} & A^{01} \\ A^{10} & A^{11} \end{bmatrix} \begin{bmatrix} \hat{s}_0 \\ \hat{s}_1 \end{bmatrix} = \hat{s}'A^{-1}\hat{s} \qquad [4]$$

and an estimating equation for σ_s^2 is (De Veer, 1986):

$$\hat{\sigma}_s^2 = [\hat{s}'A^{-1}\hat{s} + \hat{\sigma}_e^2 \text{trace}(A^{-1}C_{ss})]/(n_0 + n_1) \qquad [5]$$

where C_{ss} is the portion of the inverse of the MME in [3] corresponding to sires, and n_1 and n_0 are the number of sires with and without daughters, respectively. Many sires without daughters are often included in the sire vector and a large matrix must be inverted in each round of iteration to obtain

C_{ss}.

A possibly less costly approach would be to use a quadratic which is a function of only sires with daughters. At convergence, the estimates of sire variance from this quadratic or a quadratic including all sires should be the same because the variance structure of the records is the same. The proof that the two quadratics are equal is adapted from De Veer (1986) who credits Dr. R.L. Quaas of Cornell University for the result.

The derivation of this alternative quadratic requires several matrix results. Following the notation of [3], it can be shown (Searle, 1982) using partitioned matrix results:

$$\mathbf{A}^{01} = -\mathbf{A}^{00}\mathbf{A}_{01}(\mathbf{A}^{11})^{-1} . \qquad [6]$$

In addition:

$$\mathbf{A} * \mathbf{A}^{-1} = \mathbf{I},$$

$$\begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} \\ \mathbf{A}_{10} & \mathbf{A}_{11} \end{bmatrix} * \begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}, \text{ and}$$

$$\begin{bmatrix} \mathbf{A}_{00}\mathbf{A}^{00}+\mathbf{A}_{01}\mathbf{A}^{10} & \mathbf{A}_{00}\mathbf{A}^{01}+\mathbf{A}_{01}\mathbf{A}^{11} \\ \mathbf{A}_{10}\mathbf{A}^{00}+\mathbf{A}_{11}\mathbf{A}^{10} & \mathbf{A}_{10}\mathbf{A}^{01}+\mathbf{A}_{11}\mathbf{A}^{11} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} . \qquad [7]$$

Transposing the lower half of [7] gives:

$$(\mathbf{A}_{10}\mathbf{A}^{00}+\mathbf{A}_{11}\mathbf{A}^{10})' = \mathbf{A}^{00}\mathbf{A}_{01}+\mathbf{A}^{01}\mathbf{A}_{11} = \mathbf{0} \text{ and} \qquad [7a]$$

$$(\mathbf{A}_{10}\mathbf{A}^{01}+\mathbf{A}_{11}\mathbf{A}^{11})' = \mathbf{A}^{10}\mathbf{A}_{01}+\mathbf{A}^{11}\mathbf{A}_{11} = \mathbf{I} . \qquad [7b]$$

The first step of the derivation is to express the vector

of sires without daughters, s_0, as a function of sires with daughters, s_1. From the second equation of the MME in [3],

$$\mathbf{A}^{00}\alpha\hat{s}_0 + \mathbf{A}^{01}\alpha\hat{s}_1 = \mathbf{0}$$

solve for $\hat{s}_0$:

$$\hat{s}_0 = -(\mathbf{A}^{00})^{-1}\mathbf{A}^{01}\hat{s}_1 \ . \qquad [8]$$

Next, substitute for $\mathbf{A}^{01}$ from [6]:

$$\hat{s}_0 = -(\mathbf{A}^{00})^{-1}[-\mathbf{A}^{00}\mathbf{A}_{01}(\mathbf{A}_{11})^{-1}]\hat{s}_1,$$

$$\hat{s}_0 = \mathbf{A}_{01}(\mathbf{A}_{11})^{-1}\hat{s}_1 \qquad [9]$$

and transpose:

$$\hat{s}_0' = \hat{s}_1'(\mathbf{A}_{11})^{-1}\mathbf{A}_{10} \ . \qquad [10]$$

The right-hand side of the quantities in [10] and [9] are substituted for $\hat{s}_0'$ and $\hat{s}_0$ in the original quadratic form [4]:

$$\begin{bmatrix} \hat{s}_0' & \hat{s}_1' \end{bmatrix}\begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix}\begin{bmatrix} \hat{s}_0 \\ \hat{s}_1 \end{bmatrix} \qquad [4]$$

$$= \begin{bmatrix} \hat{s}_1'(\mathbf{A}_{11})^{-1}\mathbf{A}_{10} & \hat{s}_1' \end{bmatrix}\begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{01}(\mathbf{A}_{11})^{-1}\hat{s}_1 \\ \hat{s}_1 \end{bmatrix} .$$

The next step is to factor out $\hat{s}_1'$ on the left and $(\mathbf{A}_{11})^{-1}\hat{s}_1$ on the right-hand side:

$$= \hat{s}_1'\begin{bmatrix} (\mathbf{A}_{11})^{-1}\mathbf{A}_{10} : \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix}\begin{bmatrix} \mathbf{A}_{01} \\ \mathbf{A}_{11} \end{bmatrix}\begin{bmatrix} (\mathbf{A}_{11})^{-1}\hat{s}_1 \end{bmatrix} .$$

Multiply the two center matrices in A:

$$= \hat{s}_1' \left[(A_{11})^{-1} A_{10} : I \right] \begin{bmatrix} A^{00}A_{01} + A^{01}A_{11} \\ A^{10}A_{01} + A^{11}A_{11} \end{bmatrix} \left[(A_{11})^{-1} \hat{s}_1 \right]$$

and substitute the null and identity matrices from [7a] and [7b] for the center matrix and multiply the matrices:

$$= \hat{s}_1' \left[(A_{11})^{-1} A_{10} : I \right] \begin{bmatrix} 0 \\ I \end{bmatrix} \left[(A_{11})^{-1} \hat{s}_1 \right]$$

$$= \hat{s}_1' \left[0+I \right] \left[(A_{11})^{-1} \hat{s}_1 \right]$$

$$= \hat{s}_1' (A_{11})^{-1} \hat{s}_1 \ . \qquad [11]$$

Therefore, the quadratic in [4] consisting of both $\hat{s}_1$ and $\hat{s}_0$, sires with and without daughters, respectively, is equivalent to the quadratic in [11] which includes only those sires with daughters. Using the quadratic in [11] an estimating equation for σ_s^2 is (De Veer, 1986):

$$\hat{\sigma}_s^2 = \{\hat{s}_1' (A_{11})^{-1} \hat{s}_1 + \hat{\sigma}_e^2 \text{trace}[(A_{11})^{-1} C_{s_1 s_1}]\}/n_1 \ . \qquad [12]$$

This equation requires $(A_{11})^{-1}$, the inverse of the portion of **A** corresponding to sires with daughters, which is not the same as A^{11} from the inverse of the full relationship matrix for s_0 and s_1. One approach for generating $(A_{11})^{-1}$ is to set up the numerator relationship matrix **A** and then invert A_{11}, the section of the matrix corresponding to sires with daughters. An alternative approach for generating $(A_{11})^{-1}$ is to set up A^{-1} for all sires and then to absorb A^{00}, the section corresponding to sires without daughters.

A numerical example will show that the two procedures yield the same $(\mathbf{A}_{11})^{-1}$ matrix. Suppose sires one and two, both who have female progeny, are paternal half-sibs, i.e., they have a common sire, s_0 which does not have any female progeny. The first method for generating $(\mathbf{A}_{11})^{-1}$ requires forming the relationship matrix for all sires. Hudson et al. (1982) have presented an algorithm which allows computation of the nonzero elements of a relatively large relationship matrix with minimum memory requirements. For the example:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} \\ \mathbf{A}_{01} & \mathbf{A}_{11} \end{bmatrix} = \begin{array}{c} \begin{matrix} s_0 & s_1 & s_2 \end{matrix} \\ \left[\begin{array}{c|cc} 1 & 1/2 & 1/2 \\ \hline 1/2 & 1 & 1/4 \\ 1/2 & 1/4 & 1 \end{array}\right] \end{array}.$$

The matrix $(\mathbf{A}_{11})^{-1}$ is then formed by inverting $\mathbf{A}_{11}$:

$$(\mathbf{A}_{11})^{-1} = \begin{bmatrix} 1 & 1/4 \\ 1/4 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 16/15 & -4/15 \\ -4/15 & 16/15 \end{bmatrix}.$$

This approach for generating $(\mathbf{A}_{11})^{-1}$ is applicable when the number of sires with daughters is small relative to the number of sires without daughters.

The second approach for generating $(\mathbf{A}_{11})^{-1}$ requires $\mathbf{A}^{-1}$, the inverse of the full relationship matrix, which can be easily formed by the rules of Henderson (1975b). For the example this matrix is:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix} = \left[\begin{array}{c:cc} 5/3 & -2/3 & -2/3 \\ \hdashline -2/3 & 4/3 & 0 \\ -2/3 & 0 & 4/3 \end{array}\right].$$

The next step is to absorb $\mathbf{A}^{00}$, the section of $\mathbf{A}^{-1}$ corresponding to sires without daughters. From results for partitioned matrices (Searle, 1982) $(\mathbf{A}_{11})^{-1}$ can be generated as:

$$(\mathbf{A}_{11})^{-1} = \mathbf{A}^{11} + \mathbf{A}^{10}(-\mathbf{A}^{00})^{-1})\mathbf{A}^{01} \qquad [13]$$

which for the example is:

$$(\mathbf{A}_{11})^{-1} = \begin{bmatrix} 4/3 & 0 \\ 0 & 4/3 \end{bmatrix} + \begin{bmatrix} -2/3 \\ -2/3 \end{bmatrix} \left[\begin{bmatrix} -5/3 \end{bmatrix}^{-1} \begin{bmatrix} -2/3 & -2/3 \end{bmatrix}\right]$$

$$= \begin{bmatrix} 16/5 & -4/5 \\ -4/5 & 16/15 \end{bmatrix}.$$

This second method for generating $(\mathbf{A}_{11})^{-1}$ is applicable when the number of sires without daughters is small relative to the number of sires with daughters [computer programs to build $(\mathbf{A}_{11})^{-1}$ by forming $\mathbf{A}^{-1}$ and absorbing $\mathbf{A}^{00}$ are described in Appendix C].

Numerical example:

Meyer (1987) presented a data set consisting of 294 progeny of five sires assigned to six treatment subclasses. Let the model include treatments (t_i) as a fixed effect, and

sires (s_j) as a random effect:

$$y_{ijk} = t_i + s_j + e_{ijk}$$

where y_{ijk} is the record for the k^{th} progeny of sire j in treatment class i, and e_{ijk} is the random residual associated with y_{ijk}. Consider the following pedigree:

Bull	Sire
s_{01}	...
s_{02}	...
s_{11}	s_{01}
s_{12}	s_{01}
s_{13}	...
s_{14}	s_{02}
s_{15}	s_{02}

Sires s_{01} and s_{02} do not have progeny with records but are the sires of s_{11} and s_{12}, and s_{14} and s_{15}, respectively. The full numerator relationship matrix is:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{00} & \mathbf{A}_{01} \\ \mathbf{A}_{10} & \mathbf{A}_{11} \end{bmatrix} = \begin{array}{c} \begin{matrix} s_{01} & s_{02} & s_{11} & s_{12} & s_{13} & s_{14} & s_{15} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 1/2 & 1/2 \\ & & 1 & 1/4 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 \\ & & \text{symmetric} & & 1 & 0 & 0 \\ & & & & & 1 & 1/4 \\ & & & & & & 1 \end{bmatrix} \end{array}$$

with inverse:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{00} & \mathbf{A}^{01} \\ \mathbf{A}^{10} & \mathbf{A}^{11} \end{bmatrix} = \left[\begin{array}{cc:ccccc} 5/3 & 0 & -2/3 & -2/3 & 0 & 0 & 0 \\ & 5/3 & 0 & 0 & 0 & -2/3 & -2/3 \\ \hdashline & & 4/3 & 0 & 0 & 0 & 0 \\ & & & 4/3 & 0 & 0 & 0 \\ & & & & 1 & 0 & 0 \\ & & & & & 4/3 & 0 \\ & & & & & & 4/3 \end{array}\right].$$

The inverse of the relationship matrix for sires with daughters is:

$$(\mathbf{A}_{11})^{-1} = \begin{bmatrix} 16/15 & -4/15 & 0 & 0 & 0 \\ & 16/15 & 0 & 0 & 0 \\ \text{symmetric} & & 1 & 0 & 0 \\ & & & 16/15 & -4/15 \\ & & & & 16/15 \end{bmatrix}.$$

The least squares equations (LSE), after absorption of treatment effects, are:

$$\begin{bmatrix} \mathbf{Z'MZ} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{Z'My} \end{bmatrix} \qquad [14]$$

where **X** and **Z** are incidence matrices for t and s, respectively, $\mathbf{M}=\mathbf{I}-\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$, and y is the observation vector. For the example data, the LSE are:

$$\mathbf{Z'MZ}=\begin{bmatrix} 41.97 & -8.03 & -9.68 & -11.24 & -13.02 \\ & 42.30 & -12.23 & -8.28 & -13.76 \\ & & 46.03 & -10.53 & -13.58 \\ & & & 41.94 & -11.87 \\ & & & & 52.22 \end{bmatrix} \text{ and } \mathbf{Z'My}=\begin{bmatrix} 91.49 \\ -115.70 \\ -221.31 \\ 116.45 \\ 129.07 \end{bmatrix}.$$

The MME for all sires are:

$$\begin{bmatrix} \mathbf{A}^{00}\alpha & \mathbf{A}^{01}\alpha \\ \mathbf{A}^{10}\alpha & \mathbf{Z'MZ}+\mathbf{A}^{11}\alpha \end{bmatrix}\begin{bmatrix} \hat{s}_0 \\ \hat{s}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{Z'My} \end{bmatrix} \qquad [15]$$

and the MME for only those sires with daughters are:

$$\begin{bmatrix} \mathbf{Z'MZ}+(\mathbf{A}_{11})^{-1}\alpha \end{bmatrix}\begin{bmatrix} \hat{s}_1 \end{bmatrix} = \begin{bmatrix} \mathbf{Z'My} \end{bmatrix} \qquad [16]$$

where $\alpha=\sigma_e^2/\sigma_s^2$. For the example data, a matrix of order $n_0+n_1=7$ for MME [15] and a matrix of order $n_1=5$ for MME [16] are inverted to estimate sire variance via equations [5] and [12], respectively. The residual variance was estimated in each round of iteration from:

$$\hat{\sigma}_e^2 = (\mathbf{y'My} - \hat{\mathbf{s}}'\mathbf{Z'My})/[N - r(\mathbf{X})] \qquad [17]$$

where N is the total number of observations and $r(\mathbf{X})$ is the rank of $\mathbf{X}$, which are 294 and 6, respectively, for the example. The initial variance estimates were $\hat{\sigma}_s^2=10.0$ and $\hat{\sigma}_e^2=120.0$ and iteration was continued until the change in sire variance was less than .0001. Estimates of sire variance by round of iteration for the two quadratics are in Figure 1.

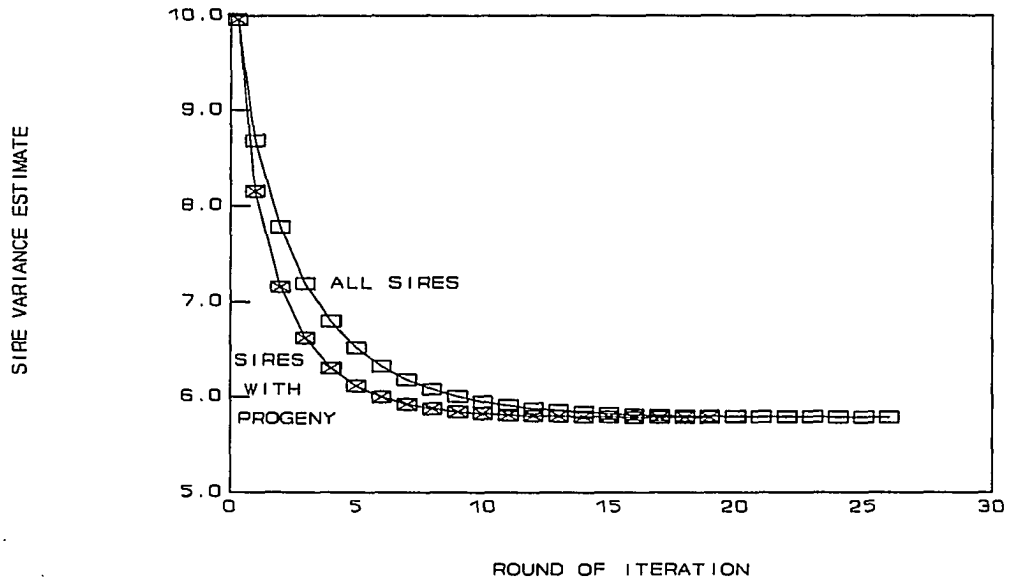

Figure 1. Estimates of sire variance by round of iteration for all sires and only sires with progeny

Although the variance estimates at convergence were the same ($\hat{\sigma}^2_s$=5.8, $\hat{\sigma}^2_e$=153.8) for both quadratics, the values at each round of iteration were not the same. The quadratic using all sires required 26 rounds and the quadratic using only sires with daughters required 19 rounds to reach convergence. These results agree with De Veer (1986) who postulated that more rounds of iteration are required to meet the convergence criterion when sires without daughters are included than when only sires with daughters are included. Therefore, absorbing equations for sires without daughters is likely to reduce computational costs by decreasing the order of the matrix to be inverted in each round of iteration and by

decreasing the number of rounds of iteration. The method can also be extended to a model with cows nested within sires. In a sire and nested cow model, cow equations are usually absorbed but the equations for sires without daughters are not affected by this absorption.

APPENDIX C. FORTRAN PROGRAMS TO BUILD THE INVERSE OF A RELATIONSHIP MATRIX DUE TO SIRES AND MATERNAL GRANDSIRES AND ABSORB SIRES WITHOUT DAUGHTERS

AINV1 and AINV2 are FORTRAN programs to build the inverse of the relationship matrix $\mathbf{A}^{-1}$ due to sires and maternal grandsires by the rules of Henderson (1975b). The section of $\mathbf{A}^{-1}$ corresponding to sires without daughters is absorbed to give a matrix of order equal to the number of bulls with daughters. This absorption reduces computational requirements of iterative variance component estimation algorithms, e.g., restricted maximum likelihood via the expectation maximization algorithm.

AINV1:

This program reads the file containing the identification number of each bull and of its sire and maternal grandsire. Bulls are renumbered consecutively one to the total number of bulls. The IDs of sires and maternal grandsires without daughters are written to an external file and a system sort in ascending order is performed. Three files are output by AINV1:

a) YSPED1 - contains one record for each bull consisting of consecutive bull ID, original bull ID, sire ID, and maternal grandsire ID (the file does not need to be sorted);

b) NBULLS - contains count of total number of bulls; and

c) SMGSFS - contains ID numbers of sires and maternal grandsires without daughters sorted in ascending order.

To use the program:

1) Pedigree information for each bull consisting of bull, sire, and maternal grandsire identification number must be sorted in ascending order and stored in FT10F001. If a sire or maternal grandsire is unknown, an ID of zero should be used for that ancestor.
2) The dimension of the vector BULLS should be equal to the number of bulls with daughters.
3) FORMAT statement 15 must be specified for reading the pedigree information file in unit 10.

AINV2:

This program reads the files created by AINV1 and expands the original bull vector (s_1) to include sires and maternal grandsires (s_0) with more than one son. $\mathbf{A}^{-1}$ is built for the expanded bull vector (s_1:s_0). The section corresponding to sires and maternal grandsires $\mathbf{A}^{00}$ is then absorbed into the bull section $\mathbf{A}^{11}$. The resulting matrix is the inverse of the relationship matrix corresponding to bulls with daughters, $(\mathbf{A}_{11})^{-1}$. This matrix is upper half-stored by consecutive rows and written to file AINVABS.

To use the program:

1) First run program AINV1 which outputs files used in AINV2.

2) The parameters NB and NSM should be set equal to the number of bulls and the number of sires and maternal grandsires, respectively. This information is included in the output from AINV1.

Numerical example:

Consider the following pedigree:

Bull	Sire	Maternal Grandsire
100	1000	0
200	2000	1000
300	100	0

Bulls 100, 200, and 300 have daughters with records and bull 100 is also the sire of bull 300. Sire 1000 does not have daughters with records but is the sire and maternal grandsire of bulls 100 and 200, respectively. Sire 2000 has only one relative so it does not need to be included in the relationship matrix. The maternal grandsires of bulls 100 and 300 are unknown so they are assigned an ID of zero. The relationship matrix $\mathbf{A}_{11}$ for bulls is:

$$\mathbf{A}_{11} = \begin{array}{c} \begin{matrix} s_{100} & s_{200} & s_{300} \end{matrix} \\ \begin{bmatrix} 1 & .125 & .5 \\ .125 & 1 & .0625 \\ .5 & .0625 & 1 \end{bmatrix} \end{array}$$

with inverse:

$$(\mathbf{A}_{11})^{-1} = \begin{bmatrix} 1.3492 & -.1270 & -.6667 \\ -.1270 & 1.0159 & 0 \\ -.6667 & 0 & 1.333 \end{bmatrix} .$$

The output from AINV1 for the example pedigree is (ID's are listed in parentheses but not printed by the program):

No. OF BULLS READ = 3 (100, 200, 300)

No. OF SIRES & MGS FOUND IN BULL LIST = 1 (100)

No. OF SIRES & MGS WRITTEN = 3 (1000, 2000, 1000).

The output from AINV2 is:

No. OF BULLS = 3 (100, 200, 300)

No. OF SIRES AND MGS READ = 3 (1000, 2000, 1000)

No. OF SIRES AND MGS ABSORBED = 1 (1000)

No. OF UNRELATED BULLS = 0

No. OF BULLS WITH SIRES FOUND = 2 (100, 300)

No. OF BULLS WITH MGS FOUND = 1 (200)

No. OF BULLS WITH SIRE & MGS FOUND = 0 .

The upper half-stored matrix written to AINVABS is:

```
 1.3492
-0.1270
-0.6667
 1.0159
 0.0000
 1.3333 .
```

The matrix built by programs AINV1 and AINV2 by setting up $\mathbf{A}^{-1}$ and then absorbing $\mathbf{A}^{00}$, the section for sires without progeny,

is the same as the inverse of the relationship matrix for only those sires with daughters.

```
//AINV1 JOB
/*JOBPARM DUPLEX=NO,FLASH=NONE,KEEP=YES
//STEP0 EXEC SCRUNC
//SYSIN DD *
K.I3446.YSPED1
K.I3446.SMGSFS
K.I3446.NBULLS
//S1 EXEC FORTVCLG,FVPOPT=2
//FORT.SYSIN DD *
C=============================================================
      PROGRAM AINV1
C=============================================================
C     PURPOSE : 1 OF 2 PROGRAMS TO READ FILE (YSPED) AND
C               OUTPUT A-1 FOR BULLS WITH SIRES AND MGS
C               ABSORBED
C
C     STEP 1: READ FILE (YSPED), RENUMBER BULLS 1 TO N OUTPUT
C             WITH SIRE,MGS TO FILE (YSPED1) AND OUTPUT SIRE,
C             MGS IDS TO FILE (SMGSF) WHICH IS THEN SORTED TO
C             FILE (SMGSFS)
C-------------------------------------------------------------
      INTEGER NBULLS,SMGSC,BULLS,HI,FIND
      DIMENSION IVEC(3),BULLS(3)
      NBULLS=0
      SMGSC=0
      N=0
      NB=0

C     loop to read thru data and read bulls into vector
  25  CONTINUE
      READ(10,15,END=9)IVEC
  15  FORMAT(I7,2I9)
      NBULLS=NBULLS+1
      BULLS(NBULLS)=IVEC(1)
      GO TO 25
   9  CONTINUE

C     loop to read thru data
      REWIND 10

  50  CONTINUE
      READ(10,15,END=99)IVEC
      NB=NB+1
C     output sire if id > 0 and not in bull list
      IF(IVEC(2) .NE. 0)THEN
C     binary search bull list vector for sire
       HI=NBULLS
       LOW=1
       FND=0

       DO 20 K=1,NBULLS
```

```
          IF(HI .LT. LOW .OR. FND .EQ. 1)GO TO 30
          MID=(HI+LOW)/2
          IF(IVEC(2) .EQ. BULLS(MID))THEN
           FND=1
           IVEC(2)=MID
           N=N+1
          ELSEIF (IVEC(2) .GT. BULLS(MID))THEN
           LOW=MID+1
          ELSE
           HI=MID-1
          END IF
 20      CONTINUE
 30      CONTINUE

         IF(FND .EQ. 0) THEN
          WRITE(20)IVEC(2)
          SMGSC=SMGSC+1
         END IF
        END IF

C       output mgs if id > 0 and not in bull list
        IF(IVEC(3) .NE. 0)THEN
C       binary search bull list vector for mgs
         HI=NBULLS
         LOW=1
         FND=0

         DO 120 K=1,NBULLS
          IF(HI .LT. LOW .OR. FND .EQ. 1)GO TO 130
          MID=(HI+LOW)/2
          IF(IVEC(3) .EQ. BULLS(MID))THEN
           FND=1
           IVEC(3)=MID
           N=N+1
          ELSEIF (IVEC(3) .GT. BULLS(MID))THEN
           LOW=MID+1
          ELSE
           HI=MID-1
          END IF
120      CONTINUE
130      CONTINUE

         IF(FND .EQ. 0) THEN
          WRITE(20)IVEC(3)
          SMGSC=SMGSC+1
         END IF
        END IF

C       output bull (numbered consecutively), sire, mgs
        WRITE(25)NB,IVEC
        GO TO 50
```

```
  99  CONTINUE

C     output results
      WRITE(30)NBULLS
      WRITE(6,*) 'No. OF BULLS READ =',NBULLS
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF SIRES & MGS FOUND IN BULL LIST =',N
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF SIRES & MGS WRITTEN =',SMGSC
      END
//GO.FT10F001 DD *
    100     1000        0
    200     3000     1000
    300      100        0
//GO.FT20F001 DD UNIT=SCRTCH,DSN=&TEMP,DISP=(NEW,PASS),
// SPACE=(TRK,(5,1),RLSE),DCB=(RECFM=FB,LRECL=4,BLKSIZE=19068)
//GO.FT25F001 DD UNIT=DISK,DSN=K.I3446.YSPED1,
// DISP=(NEW,CATLG),SPACE=(TRK,(5,1),RLSE),
// DCB=(RECFM=FB,LRECL=16,BLKSIZE=19056)
//GO.FT30F001 DD UNIT=DISK,DSN=K.I3446.NBULLS,
// DISP=(NEW,CATLG),SPACE=(TRK,(1,1),RLSE),
// DCB=(RECFM=FB,LRECL=4,BLKSIZE=19068)
//*
//STEP2 EXEC SYMSORT,TRACKS=10
//*SORTS SIRE,MGS FILE (SMGSF(&TEMP)) TO FILE (SMGSFS)
//* TRACKS = (1.3 * N * L)/19000
//SORTIN DD DSN=&TEMP,UNIT=SCRTCH,DISP=(OLD,DELETE)
//SORTOUT DD UNIT=DISK,DISP=(NEW,CATLG),DSN=K.I3446.SMGSFS,
//  SPACE=(TRK,(5,1),RLSE),DCB=(RECFM=FB,
//  LRECL=4,BLKSIZE=19068,BUFNO=1)
//SYSIN DD *
      SORT FIELDS=(1,4,BI,A)
/*
```

```
//AINV2 JOB
/*JOBPARM DUPLEX=NO,FLASH=NONE,KEEP=YES
//S0 EXEC SCRUNC
//SYSIN DD *
K.I3446.H.AINVABS
//S1 EXEC FORTVCLG,FVPOPT=2
//FORT.SYSIN DD *
C=============================================================
      PROGRAM AINV2
C=============================================================
C     PURPOSE : PROGRAM TO READ FILE (YSPED1) AND OUTPUT A-1
C               FOR BULLS WITH SIRES AND MGS ABSORBED TO FILE
C               (AINVABS)
C
C     STEPS   : READS #BULLS B FROM FILE(NBULLS); READS SORTED
C               SIRE,MGS FILE (SMGSFS) AND RENUMBERS SIRES, MGS
C               (B+1,B+2,...) IF >1 BULL;
C               READS PEDIGREE FILE (YSPED1) AND ASSIGNS
C               COEFFECIENTS TO FULL A-1 IF SIRE AND/OR MGS
C               FOUND IN LIST;
C               ABSORBS A22 (S,MGS PORTION) INTO A11 (BULL
C               PORTION) AND WRITES UPPER TRIANGULAR TO FILE
C               (AINVABS)
C
C-------------------------------------------------------------
      IMPLICIT DOUBLE PRECISION (A-H,O-Z)
      PARAMETER (NB=3,NSM=3,
     *NBSM=NB+NSM,
     *NBV=NB*(NB+1)/2,
     *NSMV=NSM*(NSM+1)/2,
     *NBSMV=NBSM*(NBSM+1)/2 )

      INTEGER OSMGS,SMGSC,SMGS,GSMGSC,SMGSV,HI,FND,SCNT,SIRES
      DIMENSION IVEC(4),SMGSV(NSM),AINV(NBSMV),AI12(NB,NSM),
      *AI22(NSMV),AINVA(NBV),IFLAG(NSM),WORK(NSM),
     *AI2221(NSM,NB)

      GSMGSC=0
      SMGSC=0
      OSMGS=0
      CNT=0
      A1=1.D0/11.D0
      A2=A1+A1
      A3=-(A2+A2)
      A4=-A3
      A5=-(A4+A4)
      A6=-(A5+A5)
      A7=1.D0/3.D0
      A8=-(A7+A7)
      A9=-(A8+A8)
      A10=1.D0/15.D0
```

```
      A11=-4.D0*A10
      A12=-4.D0*A11

      READ(30)NBULLS
C     loop to read thru s,mgs data and renumber s,mgs
C     with >=2 sons b+1,b+2,....
   50 CONTINUE
      READ(40,END=99)SMGS
      IF(SMGS .EQ. 0)GO TO 50
      SMGSC=SMGSC+1
C     new s,mgs
      IF (SMGS .NE. OSMGS)THEN
       IF(CNT .GT. 1)THEN
        GSMGSC=GSMGSC+1
        SMGSV(GSMGSC)=OSMGS
       END IF
        OSMGS=SMGS
        CNT=1
      ELSE
C     same s,mgs
       CNT=CNT+1
      END IF

      GO TO 50

   99 CONTINUE
C     check last record
      IF(CNT .GT. 1)THEN
       GSMGSC=GSMGSC+1
       SMGSV(GSMGSC)=OSMGS
      END IF

      N=NBULLS+GSMGSC

C     loop to read thru pedigree data

  100 CONTINUE
      READ(25,END=199)IVEC
      SCNT=SCNT+1
C     binary search gsmgsv for s
       IF(IVEC(3) .EQ. 0 .OR. IVEC(3) .LE. NBULLS)GO TO 135
       HI=GSMGSC
       LOW=1
       FND=0

       DO 20 K=1,GSMGSC
        IF(HI .LT. LOW .OR. FND .EQ. 1)GO TO 30
        MID=(HI+LOW)/2
        IF(IVEC(3) .EQ. SMGSV(MID))THEN
         FND=1
         IVEC(3)=MID+NBULLS
```

```
         ELSEIF (IVEC(3) .GT. SMGSV(MID))THEN
          LOW=MID+1
         ELSE
          HI=MID-1
         END IF
'20     CONTINUE
 30     CONTINUE
        IF(FND .EQ. 0)THEN
C        WRITE(6,*)IVEC(3),'NOT FOUND'
         IVEC(3)=0
       END IF

 135   CONTINUE
C      binary search gsmgsv for mgs
        IF(IVEC(4) .EQ. 0 .OR. IVEC(4) .LE. NBULLS)GO TO 150
        HI=GSMGSC
        LOW=1
        FND=0

        DO 120 K=1,GSMGSC
         IF(HI .LT. LOW .OR. FND .EQ. 1)GO TO 130
         MID=(HI+LOW)/2
         IF(IVEC(4) .EQ. SMGSV(MID))THEN
          FND=1
          IVEC(4)=MID+NBULLS
         ELSEIF (IVEC(4) .GT. SMGSV(MID))THEN
          LOW=MID+1
         ELSE
          HI=MID-1
         END IF
 120    CONTINUE
 130    CONTINUE
        IF(FND .EQ. 0)THEN
C        WRITE(6,*)IVEC(4),'NOT FOUND'
         IVEC(4)=0
       END IF
 150   CONTINUE

C      add coeffecients to appropriate row and column of A-1

       I=IVEC(1)
       J=IVEC(3)
       M=IVEC(4)

C      sire unknown, mgs unknown
       IF(J .EQ. 0 .AND. M .EQ. 0)THEN
        AINV(IHMSSF(I,I,N))=AINV(IHMSSF(I,I,N))+1.D0

C      sire unknown, mgs known
       ELSEIF(J .EQ. 0)THEN
        AINV(IHMSSF(M,M,N))=AINV(IHMSSF(M,M,N))+A10
```

```
        AINV(IHMSSF(I,M,N))=AINV(IHMSSF(I,M,N))+A11
        AINV(IHMSSF(I,I,N))=AINV(IHMSSF(I,I,N))+A12
        MGS=MGS+1

C       sire known, mgs unknown
       ELSEIF(M .EQ. 0)THEN
        AINV(IHMSSF(J,J,N))=AINV(IHMSSF(J,J,N))+A7
        AINV(IHMSSF(I,J,N))=AINV(IHMSSF(I,J,N))+A8
        AINV(IHMSSF(I,I,N))=AINV(IHMSSF(I,I,N))+A9
        SIRES=SIRES+1

C       sire known, mgs known
       ELSE
        AINV(IHMSSF(M,M,N))=AINV(IHMSSF(M,M,N))+A1
        AINV(IHMSSF(J,M,N))=AINV(IHMSSF(J,M,N))+A2
        AINV(IHMSSF(I,M,N))=AINV(IHMSSF(I,M,N))+A3
        AINV(IHMSSF(J,J,N))=AINV(IHMSSF(J,J,N))+A4
        AINV(IHMSSF(I,J,N))=AINV(IHMSSF(I,J,N))+A5
        AINV(IHMSSF(I,I,N))=AINV(IHMSSF(I,I,N))+A6
        MGS=MGS+1
        SIRES=SIRES+1
       END IF

       GO TO 100

  199 CONTINUE

C      add 1 to diagonals corresponding to sires,mgs
       DO 299 I=(NBULLS+1),N
       AINV(IHMSSF(I,I,N))=AINV(IHMSSF(I,I,N))+1.D0
 299   CONTINUE
1000   CONTINUE
C     absorb sires, mgs into bulls

C      assign coef. to fullstored (row*col) AI12
       DO 300 I=1,NBULLS
        DO 300 J=1,GSMGSC
        AI12(I,J)=AINV(IHMSSF(I,(J+NBULLS),N))
 300   CONTINUE


C      assign coef. to halfstored AI22
       L=(NBULLS*(NBULLS+1)/2)+(GSMGSC*NBULLS)
       DO 400 I=(NBULLS+1),N
        DO 400 J=I,N
         K=IHMSSF(I,J,N)
        AI22(K-L)=AINV(K)
 400   CONTINUE

C      invert A22 for absorption
       CALL DKMVHF(AI22,WORK,IFLAG,GSMGSC)
```

```
C     multiply A12*((A22)-1)*A12'

C     postmultiply AI22 with A12', store in AI2221
      N1=GSMGSC+1
      DO 5 J=1,NBULLS
       DO 5 I=1,GSMGSC
        XX=0.D0
        IK=I-GSMGSC
        DO 6 K=1,GSMGSC
         CALL IROWHF(IK,I,K,N1)
 6       XX=XX+AI22(IK)*AI12(J,K)
 5    AI2221(I,J)=XX

C     premultiply AI2221 with AI12, store in AINVA
       IJ=0
       DO 7 I=1,NBULLS
        DO 7 J=I,NBULLS
         IJ=IJ+1
         XX=0.D0
         DO 8 K=1,GSMGSC
  8       XX=XX+AI12(I,K)*AI2221(K,J)
  7    AINVA(IJ)=XX

C     subtract A11-[A12*((A22)-1)*A12']
      K=0
      N=NBULLS+GSMGSC
      DO 600 I=1,NBULLS
       DO 600 J=I,NBULLS
        K=K+1
        AINVA(K)=AINV(IHMSSF(I,J,N))-AINVA(K)
C       WRITE(50)AINVA(K)
        WRITE(6,62)AINVA(K)
 62   FORMAT(F9.4,5X)
 600  CONTINUE

C     output results
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF SIRES AND MGS READ',SMGSC
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF SIRES AND MGS ABSORBED',GSMGSC
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF BULLS WITH SIRES FOUND',SIRES
      WRITE(6,*)' '
      WRITE(6,*) 'No. OF BULLS WITH MGS FOUND',MGS
      END
C=============================================================
      INTEGER FUNCTION IHMSSF(I,J,N)
C=============================================================
C     FUNCTION TO WORK OUT ADDRESS IN A HALFSTORED SYMMETRIC
C     MATRIX OF ORDER N; CONSIDER THE UPPER TRIANGLE
```

```
C      (I=ROW,J=COLUMN)
C=====================================================================
       IF(I.LE.J)THEN
       I1=I-1
       IHMSSF=N*I1-I*I1/2+J
       ELSE
       J1=J-1
       IHMSSF=N*J1-J*J1/2+I
       END IF
       RETURN
       END

C=====================================================================
       SUBROUTINE IROWHF(IJ,I,J,N1)
C=====================================================================
C      ROUTINE TO WORK OUT THE ADDRESS OF THE J-TH ELEMENT IN
C      THE I-TH ROW/COLUMN OF A SYMMETRIC HALFSTORED MATRIX OF
C      ORDER N WHEN ACCESSING ALL N ELEMENTS OF THE ROW IN
C      ORDER 1 TO N
C      PARAMETER SETTING REQUIRED BEFORE STARTING :
C      N1=N+1
C      IJ=I-N
C                                              KM 3/85
C---------------------------------------------------------------------
       IF(J.LE.I)THEN
       IJ=IJ+N1-J
       ELSE
       IJ=IJ+1
       END IF

       RETURN
       END
C=====================================================================
                  SUBROUTINE DKMVHF(A,V,IFLAG,N)
C=====================================================================
C  * * * ROUTINE TO INVERT A HALFSTORED SYMMETRIC MATRIX  * *
C       IF THE MATRIX IS NOT OF FULL RANK THE GENERALISED
C       INVERSE IS RETURNED, SETTING N-RANK(A) ROWS/COLUMNS TO
C       ZERO AND OBTAINING THE REGULAR INVERSE OF THE FULL RANK
C       SUBMATRIX
C       THIS IS A REWRITE OF HENDERSON'S MATRIX INVERTOR
C       "DJNVHF", USING HIS ALGORITHM BUT AVOIDING TO REARRANGE
C       ROWS AND COL.S
C       PARAMETERS :
C       A : DOUBLE PRECISION VECTOR OF LENGTH N*(N+1)/2,
C       CONTAINING THE MATRIX TO BE INVERTED HALFSTORED ON
C       ENTRY AND THE INVERSE ON EXIT
C       V : DOUBLE PRECISION VECTOR OF LENGTH N, USED AS
C       WORKSPACE
C       IFLAG : INTEGER VECTOR OF LENGTH N, CONTAINING THE
C       ORDER IN WHICH ROWS/COLS WERE PROCESSED ON EXIT, EXCEPT
```

```
C       FOR THE N-TH ELEMENT WHICH GIVES THE RANK OF THE MATRIX
C       N : ORDER OF THE MATRIX
C                                                  KARIN MEYER
C                                                  NOVEMBER 1983
C-----------------------------------------------------------------
      DOUBLE PRECISION A(1),V(1),XX,DMAX,AMAX,BMAX,ZERO,DIMAX
      INTEGER IFLAG(1)

      IF(N.EQ.1)THEN
      XX=A(1)
      IF(DABS(XX).GT.ZERO)THEN
      A(1)=1.D0/XX
      IFLAG(1)=1
      ELSE
      A(1)=0.D0
      IFLAG(1)=0
      END IF
      RETURN
      END IF

      N1=N+1
      NN=N*N1/2
      DO 1 I=1,N
    1 IFLAG(I)=0

C       SET MINIMUM ABSOLUTE VALUE OF DIAGONAL ELEMENTS FOR
C       NON-SINGULARITY (MACHINE SPECIFIC )
        ZERO=1.D-20

C-----------------------------------------------------------------
C       START LOOP OVER ROWS/COLS
C-----------------------------------------------------------------
      DO 8 II=1,N

C       ... FIND DIAGONAL ELEMENT WITH BIGGEST ABSOLUTE VALUE
           DMAX=0.D0
           AMAX=0.D0
           KK=-N
           DO 2 I=1,N
C        ... CHECK THAT THIS ROW/COL HAS NOT BEEN PROCESSED
                IF(IFLAG(I).NE.0)THEN
                KK=KK+N1-I
                ELSE
                KK=KK+N1
                BMAX=DABS(A(KK))
                IF(BMAX.GT.AMAX)THEN
              DMAX=A(KK)
                   AMAX=BMAX
                   IMAX=I
                END IF
             KK=KK-I
```

```
            END IF
    2       CONTINUE
C         ... CHECK FOR SINGULARITY
            IF(AMAX.LE.ZERO)GO TO 11
C         ... ALL ELEMENTS SCANNED,SET FLAG
            IFLAG(IMAX)=II

C         ... INVERT DIAGONAL
            DIMAX=1.D0/DMAX
C         ... DEVIDE ELEMENTS IN ROW/COL PERTAINING TO THE
C        BIGGEST DIAGONAL ELEMENT BY DMAX
            IL=IMAX-N
            DO 3 I=1,IMAX-1
                IL=IL+N1-I
                XX=A(IL)
                IF(XX.NE.0)A(IL)=XX*DIMAX
    3       V(I)=XX
C         ... NEW DIAGONAL ELEMENT
            IL=IL+N1-IMAX
            A(IL)=-DIMAX
            DO 4 I=IMAX+1,N
                IL=IL+1
                XX=A(IL)
                IF(XX.NE.0)A(IL)=XX*DIMAX
    4       V(I)=XX
C         ... ADJUST THE OTHER ROWS/COLS :
C        A(I,J)=A(I,J)-A(I,IMAX)*A(J,IMAX)/A(IMAX,IMAX)
            IJ=0
            DO 7 I=1,N
                IF(I.EQ.IMAX)THEN
      IJ=IJ+N1-I
      ELSE
                XX=V(I)
                IF(XX.NE.0.D0)THEN
                    XX=XX*DIMAX
                    DO 5 J=I,N
                        IJ=IJ+1
                        IF(J.NE.IMAX)A(IJ)=A(IJ)-XX*V(J)
    5               CONTINUE
                ELSE
    6               IJ=IJ+N1-I
                END IF
      END IF
    7       CONTINUE

C         ... REPEAT UNTIL ALL ROWS/COLS ARE PROCESSED
    8   CONTINUE

C-------------------------------------------------------------
C        END LOOP OVER ROWS/COLS
C-------------------------------------------------------------
```

```
C       ... REVERSE SIGN
        DO 9 I=1,NN
     9  A(I)=-A(I)
C       ... AND THAT'S IT
C       PRINT 10,N
    10  FORMAT(' FULL RANK MATRIX INVERTED, ORDER =',I5)
C        RETURN RANK AS LAST ELEMENT OF FLAG VECTOR
         IFLAG(N)=N

        RETURN

C-------------------------------------------------------------
C       MATRIX NOT OF FULL RANK, RETURN GENERALISED INVERSE
C-------------------------------------------------------------
    11  IRANK=II-1
        IJ=0
        DO 14 I=1,N
            IF(IFLAG(I).EQ.0)THEN
C        ... SET REMAINING N-II ROWS/COLS TO ZERO
                DO 12 J=I,N
                    IJ=IJ+1
                    A(IJ)=0.D0
    12          CONTINUE
            ELSE
                DO 13 J=I,N
                    IJ=IJ+1
                    IF(IFLAG(J).NE.0)THEN
C     ... REVERSE SIGN FOR II-1 ROWS/COLS PREVIOUSLY PROCESSED
                        A(IJ)=-A(IJ)
                    ELSE
                        A(IJ)=0.D0
                    END IF
    13          CONTINUE
            END IF

    14  CONTINUE
        PRINT 15,N,IRANK
    15  FORMAT(' GENERALISED INVERSE OF MATRIX WITH ORDER =',I5,

      1  '   AND RANK =',I5)
        IFLAG(N)=IRANK

        RETURN
        END
//GO.FT40F001 DD UNIT=DISK,DSN=K.I3446.SMGSFS,DISP=(OLD,KEEP)
//GO.FT25F001 DD UNIT=DISK,DSN=K.I3446.YSPED1,DISP=(OLD,KEEP)
//GO.FT50F001 DD UNIT=DISK,DSN=K.I3446.AINVABS,
// DISP=(NEW,CATLG),SPACE=(TRK,(50,5),RLSE),
// DCB=(RECFM=FB,LRECL=8,BLKSIZE=19064)
//GO.FT30F001 DD UNIT=DISK,DSN=K.I3446.NBULLS,DISP=(OLD,KEEP)
```

APPENDIX D. ALTERNATIVE MODELS FOR DAIRY COW AND SIRE EVALUATION WITH REPEATED LACTATIONS

The model chosen for estimation of variance components or prediction of breeding values is usually a compromise between an ideal model and computational requirements. Simplifying assumptions are often made to reduce the computing time or memory required. The sire and nested cow model is an approximation to the animal model (AM) which ignores mates of sires and relationships through females but is less computationally demanding.

The yield of a cow, ignoring any fixed effects, can be represented with an AM as

$$y = (.5a_s + .5a_d + \emptyset) + p + e \qquad [1]$$

where the quantity in parentheses is the breeding value of the cow and is expressed as the sum of one-half the breeding values of the cow's sire (a_s) and dam (a_d) plus the Mendelian sampling effect ($\emptyset$), p is a permanent environmental effect which includes environmental and nonadditive genetic effects common to each record of the cow, and e is a temporary environmental effect. This form of the animal model allows prediction of breeding values for all individuals and producing abilities for individuals with records and also allows estimation of additive genetic, σ_a^2, permanent environmental, σ_p^2, and residual, σ_e^2, variances. The full AM requires an equation for each individual and an additional equation for each cow with a record. Estimation of variance

components in a repeated-lactation animal model via a restricted maximum likelihood (REML) algorithm which requires inversion of mixed model equations is limited to small data sets.

An alternative approach is to use an approximate animal model which ignores mates of sires and female relationships, i.e., a sire and cow nested within sire model. In the sire and nested cow model, $.5a_s$ is expressed as a separate sire effect and $.5a_d$, $\emptyset$, and p are combined in a nested cow effect:

$$y = .5a_s + (.5a_d + \emptyset + p) + e \ . \qquad [2]$$

The equations corresponding to p are eliminated so the number of equations is equal to the number of individuals. The model allows prediction of sire and cow breeding values, cow producing abilities, and estimation of sire (σ_s^2), cow (σ_c^2), and residual variance.

Henderson (1985) defined linear equivalent models as models which yield identical first and second moments of the data vector. Henderson showed that two models were linearly equivalent by demonstrating that, for a simple numerical example, the estimators and predictors obtained from one model could be converted by a linear transformation to estimators and predictors from the other model. A small numerical example will be used to show the equivalence of solutions from a sire and nested cow model and solutions from an animal model ignoring mates and female relationships.

The following records are available for the evaluation:

Cow	Sire	Records
3	1	1990, 2081, 2084
4	1	1255, 1252, 1214
5	2	1689, 1548

Sires 1 and 2 do not have records but are evaluated through their daughters. Mates of the sires (dams of the cows) are not identified. For the example, heritability (h^2) is .25 and repeatability (r) is .53125.

Sire and nested cow model:

The assumed model consists of a fixed overall mean (μ) and random sire (s), cow (c), and residual (e) effects:

$$y_{ijk} = \mu + s_i + c_{ij} + e_{ijk} \qquad [1]$$

with $E(y_{ijk})=\mu$, $E(s_i)=E(c_{ij})=E(e_{ijk})$, and

$$\text{Var}\begin{bmatrix} s \\ c \\ e \end{bmatrix} = \begin{bmatrix} I\sigma_s^2 & 0 & 0 \\ 0 & I\sigma_c^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{bmatrix}$$

where y_{ijk} is yield k of cow j of sire i, and σ_s^2, σ_c^2, and σ_e^2 are variances of s, c, and e effects, respectively. The mixed model equations (MME) are:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'W} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{I}\alpha_s & \mathbf{Z'W} \\ \mathbf{W'X} & \mathbf{W'Z} & \mathbf{W'W}+\mathbf{I}\alpha_c \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s} \\ \hat{c} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{W'y} \end{bmatrix}$$

where **X**, **Z**, and **W** are design matrices for μ, s, and c effects respectively, y is the vector of records, $\alpha_s=\sigma_e^2/\sigma_s^2=(1-r)/(.25h^2)$, and $\alpha_c=\sigma_e^2/\sigma_c^2=(1-r)/(r-.25h^2)$ (see Appendix A for a derivation of the variance ratios expressed in terms of heritabilty and repeatability), h^2 is heritability, and r is repeatability. For the numerical example, $\alpha_s=7.5$, $\alpha_c=1.0$, and the MME are:

$$\begin{bmatrix} 8 & 6 & 2 & 3 & 3 & 2 \\ & 6+7.5 & 0 & 3 & 3 & 0 \\ & & 2+7.5 & 0 & 0 & 2 \\ & \text{symmetric} & & 3+1 & 0 & 0 \\ & & & & 3+1 & 0 \\ & & & & & 2+1 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{s}_1 \\ \hat{s}_2 \\ \hat{c}_{11} \\ \hat{c}_{12} \\ \hat{c}_{21} \end{bmatrix} = \begin{bmatrix} 13113 \\ 9876 \\ 3237 \\ 6155 \\ 3721 \\ 3237 \end{bmatrix}$$

with solutions: $\hat{\mu} = 1637.0$,

$\hat{s}_i = (\ 1.5 \quad -1.5\ )$, and

$\hat{c}_{ij} = (\ 309.9 \quad -298.6 \quad -11.3\ )$.

Sire solutions predict transmitting ability or one-half of breeding value. Predicted breeding values are 3.0 and -3.0 for sires one and two, respectively. Cow solutions contain both genetic and permanent environmental effects. Because sire

effects contribute one-quarter of the total genetic variance, cow effects contain the other three-quarters of genetic variance. The fraction of a cow solution that is expected to be genetic is $(.75h^2)/(r-.25h^2)$ which is 0.4 for the example. Predicted breeding values and producing abilities for cows are:

Cow	Breeding value ($\hat{s}_i+.4\hat{c}_{ij}$)	Producing ability ($\hat{s}_i+\hat{c}_{ij}$)
3	125.5	311.4
4	-117.9	-297.1
5	-6.0	-12.8

Approximate animal model:

The assumed model consists of a fixed overall mean (μ) and random animal (a), permanent environmental (p), and residual effects:

$$y_{ij} = \mu + a_i + p_i + e_{ij}$$

with $E(y_{ijk})=\mu$, $E(a_i)=E(p_i)=E(e_{ij})=0$, and

$$\text{Var}\begin{bmatrix} a \\ p \\ e \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

where y_{ij} is yield j of animal i, **A** is the numerator relationship matrix, and σ_a^2, σ_p^2, and σ_e^2 are variances of a, p, and e effects, respectively. The MME are:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z}_a & \mathbf{X'Z}_p \\ \mathbf{Z}_a'\mathbf{X} & \mathbf{Z}_a'\mathbf{Z}_a+\mathbf{A}^{-1}\alpha_a & \mathbf{Z}_a'\mathbf{Z}_p \\ \mathbf{Z}_p'\mathbf{X} & \mathbf{Z}_p'\mathbf{Z}_a & \mathbf{Z}_p'\mathbf{Z}_p+\mathbf{I}\alpha_p \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a} \\ \hat{p} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z}_a'\mathbf{y} \\ \mathbf{Z}_p'\mathbf{y} \end{bmatrix}$$

where $\mathbf{X}$, $\mathbf{Z}_a$, and $\mathbf{Z}_p$ are design matrices for μ, a, and p effects, $\alpha_a=\sigma_e^2/\sigma_a^2$, and $\alpha_p=\sigma_e^2/\sigma_p^2$. Similar to the sire and nested cow model, the variance ratios required in the MME for the approximate animal model can be expressed as a function of heritability and repeatability. Total variance, σ_y^2, is the sum of animal, permanent environmental, and residual variances, i.e., $\sigma_y^2=\sigma_a^2+\sigma_p^2+\sigma_e^2$. Heritability is the ratio of animal to total variance, $h^2=\sigma_a^2/\sigma_y^2$, and repeatability is the ratio of the sum of animal and permanent environmental variances to total variance, $r=(\sigma_a^2+\sigma_p^2)/\sigma_y^2$. Residual variance can be reexpressed as a function of total variance and repeatability:

$$\sigma_e^2=\sigma_y^2-(\sigma_a^2+\sigma_p^2)$$
$$\sigma_e^2=\sigma_y^2-r\sigma_y^2$$
$$\sigma_e^2=(1-r)\sigma_y^2 \ .$$

Similarly, animal variance can be expressed as a function of total variance, $\sigma_a^2=h^2\sigma_y^2$, and permanent environmental variance can be expressed as a function of total variance, heritability, and repeatability:

$$\sigma_p^2=r\sigma_y^2-\sigma_a^2$$
$$\sigma_p^2=r\sigma_y^2-h^2\sigma_y^2$$

$$\sigma_p^2=(r-h^2)\sigma_y^2 \ .$$

Therefore, the variance ratios can be expressed as $\alpha_a=\sigma_e^2/\sigma_a^2=(1-r)/h^2$ and $\alpha_p=\sigma_e^2/\sigma_p^2=(1-r)/(r-h^2)$. For the example data, $\alpha_a=1.875$, $\alpha_p=1.667$, and the numerator relationship matrix **A** ordered by animal number, is:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & .5 & .5 & 0 \\ & 1 & 0 & 0 & .5 \\ & & 1 & .25 & 0 \\ \text{symmetric} & & & 1 & 0 \\ & & & & 1 \end{bmatrix}$$

with inverse:

$$\mathbf{A}^{-1} = \begin{bmatrix} 5/3 & 0 & -2/3 & -2/3 & 0 \\ & 4/3 & 0 & 0 & -2/3 \\ \text{symmetric} & & 4/3 & 0 & 0 \\ & & & 4/3 & 0 \\ & & & & 4/3 \end{bmatrix} .$$

The MME for the sample data are:

$$\begin{bmatrix} 8 & 0 & 0 & 3 & 3 & 2 & 3 & 3 & 2 \\ & 5/3\alpha_a & 0 & -2/3\alpha_a & -2/3\alpha_a & 0 & 0 & 0 & 0 \\ & & 4/3\alpha_a & 0 & 0 & -2/3\alpha_a & 0 & 0 & 0 \\ & & & 3+4/3\alpha_a & 0 & 0 & 3 & 0 & 0 \\ & & & & 3+4/3\alpha_a & 0 & 0 & 3 & 0 \\ & & & & & 2+4/3\alpha_a & 0 & 0 & 2 \\ & & & \text{symmetric} & & & 3+\alpha_p & 0 & 0 \\ & & & & & & & 3+\alpha_p & 0 \\ & & & & & & & & 2+\alpha_p \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{p}_3 \\ \hat{p}_4 \\ \hat{p}_5 \end{bmatrix} = \begin{bmatrix} 13113 \\ 0 \\ 0 \\ 6155 \\ 3721 \\ 3237 \\ 6155 \\ 3721 \\ 3237 \end{bmatrix}$$

with solutions:

$$\hat{\mu} = 1637.0 \ ,$$

$$\hat{a}_i = (\ 3.0 \quad -3.0 \quad 125.5 \quad -117.9 \quad -6.0 \ ) \ , \text{ and}$$

$$\hat{p}_i = (\ 186.0 \quad -179.2 \quad -6.8 \ ) \ .$$

Animal solutions predict breeding values and the sum of animal and permanent environmental effect is the prediction of producing ability:

Animal	Breeding value ($\hat{a}_i$)	Producing ability ($\hat{a}_i+\hat{p}_i$)
1	3.0	...
2	-3.0	...
3	125.5	311.4
4	-117.9	-297.1
5	-6.0	-12.8

These values from the approximate animal model ignoring mates of sires and relationships through females are the same as those obtained from the sire and nested cow model but not the same as those obtained from an animal model including all relationships. A full animal model includes all relationships between animals and removes biases resulting from nonrandom mating, i.e., selection of mates. The animal model may be computationally prohibitive especially for variance component estimation in large data sets. In a sire and nested cow model, the equations for permanent environment are eliminated and cow equations are easily absorbed so computational requirements are reduced.

APPENDIX E. NUMERICAL EXAMPLE OF AN ANIMAL MODEL WITH EQUAL HERITABILITY AND UNEQUAL VARIANCES ACROSS HERDS

The mixed model equations (MME) for single traits can often be obtained by simple modifications of the ordinary least squares equations. For example, in an animal model for genetic effects, the MME can be obtained by setting up the least squares equations and then adding to the animal equations the inverse of the relationship matrix, $\mathbf{A}^{-1}$, multiplied by the ratio of residual to additive genetic variance, $(\sigma_e^2/\sigma_a^2)=\alpha$. This simplified approach of forming the MME assumes that all observations have the same variance and is not applicable in the presence of heterogeneous variance across subclasses even when heritability is equal for all subclasses.

A simple numerical example adapted from Henderson (1984a) will illustrate the effect of heterogeneous variances when heritability is equal across subclasses. Consider the design:

Animal	Sire	Herd	Record
1	...	...	...
2	1	1	3
3	...	1	2
4	1	2	5
5	...	2	6

Animal 1 does not have a record but is the sire of animals 2

and 4. The numerator relationship matrix for the five animals, ordered by animal number, is:

$$\mathbf{A} = \begin{bmatrix} 1 & .5 & 0 & 5 & 0 \\ & 1 & 0 & .25 & 0 \\ & & 1 & 0 & 0 \\ \text{symmetric} & & & 1 & 0 \\ & & & & 1 \end{bmatrix}$$

with inverse:

$$\mathbf{A}^{-1} = \begin{bmatrix} 5/3 & -2/3 & 0 & -2/3 & 0 \\ & 4/3 & 0 & 0 & 0 \\ \text{symmetric} & & 1 & 0 & 0 \\ & & & 4/3 & 0 \\ & & & & 1 \end{bmatrix}.$$

The assumed model is:

$$y_{ij} = h_i + a_{ij} + e_{ij} \qquad [1]$$

where y_{ij} is the record of animal j in herd i, h_i is a fixed herd effect, a_{ij} is a random genetic effect of animal j when production is in herd i, and e_{ij} a random residual effect. Let σ_e^2 be 12 and 48 and σ_a^2 be 4 and 16 for herds one and two, respectively. A genetic correlation of 1 is assumed across herds and α=3 in each herd. Variances are greater in herd two but heritability, $h^2=\sigma_a^2/(\sigma_a^2+\sigma_e^2)$, is .25 in each herd.

Genetic variance is made equal across herds by rewriting model [1] as (Gianola, 1986):

$$y_{ij}/\sigma_{a_i} = (h_i + a_{ij} + e_{ij})/\sigma_{a_i}$$
$$y^*_{ij} = h^*_i + a^*_j + e^*_{ij} \quad . \qquad [2]$$

Each element of the original model is divided by the appropriate genetic standard deviation, i.e., 2 and 4 in herds one and two of the example. In model [2], the subscript i does not appear in the animal effect because σ^2_a=1 for both herds:

$$\begin{aligned} \text{var}(a^*_j) &= \text{var}(a_{ij}/\sigma_{a_i}) \\ &= (1/\sigma_{a_i})^2 \ \text{var}(a_{ij}) \\ &= (1/\sigma^2_{a_i}) \ \sigma^2_{a_i} \\ &= 1 \ . \end{aligned}$$

The genetic variance matrix for the transformed model is equal to **A**. The variance of the transformed residual for a record in herd i is:

$$\begin{aligned} \text{var}(e^*_{ij}) &= \text{var}(e_{ij}/\sigma_{a_i}) \\ &= (1/\sigma_{a_i})^2 \ \text{var}(e_{ij}) \\ &= (1/\sigma^2_{a_i}) \ \sigma^2_{e_i} \\ &= \sigma^2_{e_i}/\sigma^2_{a_i} \ . \end{aligned}$$

The matrix of transformed residuals $\mathbf{R}^*$ is diagonal with elements $(\sigma^2_{e_i}/\sigma^2_{a_i})=(1-h^2)/h^2$ for observations in herd i and is $\mathbf{I}*3$ for the numerical example. The transformed mixed model equations for model [2] are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{*-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{*-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{*-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{*-1}\mathbf{Z}+\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{h}^* \\ \hat{a}^* \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{*-1}y^* \\ \mathbf{Z}'\mathbf{R}^{*-1}y^* \end{bmatrix} \qquad [3]$$

where $\mathbf{X}$ and $\mathbf{Z}$ are design matrices for $\hat{h}^*$ and $\hat{a}^*$, respectively, and y^* is the transformed observation vector with elements y_{ij}/σ_{a_i}.

For the numerical example the equations from [3] are:

$$\begin{bmatrix} 2/3 & 0 & 0 & 1/3 & 1/3 & 0 & 0 \\ & 2/3 & 0 & 0 & 0 & 1/3 & 1/3 \\ & & 0+5/3 & 0-2/3 & 0 & 0-2/3 & 0 \\ & \text{symmetric} & & 1/3+4/3 & 0 & 0 & 0 \\ & & & & 1/3+1 & 0 & 0 \\ & & & & & 1/3+4/3 & 0 \\ & & & & & & 1/3+1 \end{bmatrix} \begin{bmatrix} \hat{h}_1^* \\ \hat{h}_2^* \\ \hat{a}_1^* \\ \hat{a}_2^* \\ \hat{a}_3^* \\ \hat{a}_4^* \\ \hat{a}_5^* \end{bmatrix} = \begin{bmatrix} (5/2)/3 \\ (11/4)/3 \\ 0 \\ (3/2)/3 \\ (2/2)/3 \\ (5/4)/3 \\ (6/4)/3 \end{bmatrix}$$

with solutions:

$$\hat{h}_i^* = (\ 1.254 \quad 1.367 \ ) \text{ and}$$

$$\hat{a}_j^* = (\ .015 \quad .055 \quad -.063 \quad -.017 \quad .033 \ ) .$$

To convert the solutions from the transformed model to those that would be obtained from a multiple trait model in which production in each herd is considered a different trait, the transformed solutions are multiplied by the value of σ_{a_i} corresponding to herd i (Gianola, 1986), i.e., 2 and 4 in herds one and two, respectively:

$$\hat{h}_i = (\ 2.508 \quad 5.468 \ ) ,$$

$$\hat{a}_{1j} = (\ .030 \quad .110 \quad -.127 \quad -.035 \quad .066 \ ) \ , \text{ and}$$

$$\hat{a}_{2j} = (\ .061 \quad .221 \quad -.254 \quad -.069 \quad .133 \ ) \ .$$

Therefore, $\hat{a}$ in herd two is twice $\hat{a}$ in herd one, but the animals rank the same in both herds, $a_2 > a_5 > a_1 > a_4 > a_3$, as expected from a genetic correlation of one across herds.

Consider the consequence of assuming variances are equal for the two herds. The simplified MME are the least squares equations with $\mathbf{A}^{-1}\alpha$ added to the animal equations:

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{h} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [4]$$

and for the numerical example are:

$$\begin{bmatrix} 2 & 0 & 0 & 1 & 1 & 0 & 0 \\ & 2 & 0 & 0 & 0 & 1 & 1 \\ & & 0+(5/3)3 & (-2/3)3 & 0 & (-2/3)3 & 0 \\ & & & 1+(4/3)3 & 0 & 0 & 0 \\ & & \text{symmetric} & & 1+(1)3 & 0 & 0 \\ & & & & & 1+(4/3)3 & 0 \\ & & & & & & 1+(1)3 \end{bmatrix} \begin{bmatrix} \hat{h}_1 \\ \hat{h}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix} = \begin{bmatrix} 5 \\ 11 \\ 0 \\ 3 \\ 2 \\ 5 \\ 6 \end{bmatrix} .$$

Solutions for the simplified equations are:

$$\hat{h}_i = (\ 2.516 \quad 5.484 \ ) \text{ and}$$

$$\hat{a}_j = (\ 0 \quad .097 \quad -.129 \quad -.097 \quad .129 \ )$$

and the animal rank is $a_5 > a_2 > a_1 > a_4 > a_3$. In contrast to the previous evaluation which correctly accounted for heterogeneous variances, an evaluation assuming homogeneous

variances underevaluated animal 2 in the low variance herd and overevaluated animal 5 in the high variance herd. Therefore, assuming variances are homogeneous across herds when they are actually heterogeneous can result in misranking of animals even if heritability is equal for all herds.

The method of Gianola (1986) is a general approach to account for heterogeneity of variances and can be used when heritability is not equal across herds provided that the genetic correlation across herds is equal to one. If heritability is equal across herds, however, the approach can be simplified so that estimates of genetic and residual components of variance are not required. Instead, only the ratio of total variance across herds needs to be estimated. In the numerical example the total variance ($\sigma_a^2+\sigma_e^2$) is 16 and 64 in herds one and two, respectively, resulting in a 1:4 ratio of total variances and 1:2 ratio of standard deviations. The heterogeneous variances in the two herds can be accounted for by dividing the observations in the simplified MME in [4] by the ratio of the standard deviations, i.e., 1 and 2 in herds one and two, respectively. For the numerical example the MME are:

$$\begin{bmatrix} 2 & 0 & 0 & 1 & 1 & 0 & 0 \\ & 2 & 0 & 0 & 0 & 1 & 1 \\ & & 0+(5/3)3 & (-2/3)3 & 0 & (-2/3)3 & 0 \\ & & & 1+(4/3)3 & 0 & 0 & 0 \\ & & \text{symmetric} & & 1+(1)3 & 0 & 0 \\ & & & & & 1+(4/3)3 & 0 \\ & & & & & & 1+(1)3 \end{bmatrix} \begin{bmatrix} \hat{h}_1^\bullet \\ \hat{h}_2^\bullet \\ \hat{a}_1^\bullet \\ \hat{a}_2^\bullet \\ \hat{a}_3^\bullet \\ \hat{a}_4^\bullet \\ \hat{a}_5^\bullet \end{bmatrix} = \begin{bmatrix} 5/1 \\ 11/2 \\ 0 \\ 3/1 \\ 2/1 \\ 5/2 \\ 6/2 \end{bmatrix}$$

with solutions:

$$\hat{h}_i^\bullet = (\ 2.508 \quad 2.734 \ ) \text{ and}$$

$$\hat{a}_j^\bullet = (\ .030 \quad .110 \quad -.127 \quad -.035 \quad .066 \ ) \ .$$

The multiple trait model solutions can be obtained by multiplying the above solutions by the ratio of the standard deviations, i.e., 1 and 2 in herds one and two, respectively:

$$\hat{h}_i = (\ 2.508 \quad 5.468 \ ) \ ,$$

$$\hat{a}_{1j} = (\ .030 \quad .110 \quad -.127 \quad -.035 \quad .066 \ ) \text{ , and}$$

$$\hat{a}_{2j} = (\ .061 \quad .221 \quad -.254 \quad -.069 \quad .133 \ ) \ .$$

This method to account for heterogeneous variances when heritability is equal across herds is simpler than the previous method because only estimates of the ratio of total variance across herds are required instead of estimates of genetic and residual components of variance. A practical and computationally feasible approach would be to estimate phenotypic variance within herds and then to divide the observations by the ratios of the estimated standard deviations.